

A time for responsibility

Researchers have good reasons to be nervous of proposals that all life science research grant applications to Brussels be scrutinized for their potential implications. But careful handling could produce benefits for all.

There is, at times, something almost Kafkaesque about the way in which the scientific community has recently found itself in the dock, accused of vaguely worded crimes, unable fully to comprehend the nature of the charges it faces, yet threatened with dire sanctions if found guilty by public acclaim. Such, at least, tends to be the feeling generated in researchers who find themselves bemused by the growing desire of both governments and elected politicians to embrace the discipline of bioethics. It is certainly the reaction already provoked in some by the news that the European Commission is likely to require all grant applications in the life sciences to be assessed for their moral acceptability (see page 433).

The dangers in such moves are clear. The most obvious lie in the extra burdens imposed by any new bureaucratic requirement. But requiring all applications to pass an ethics test could in principle impose a stifling moral and intellectual orthodoxy on research, pre-judging what can and cannot be studied in potentially dangerous ways. Anyone claiming that this is an exaggeration needs only to look at the powerful influence of the anti-abortion lobby on critical areas of biomedical research in the United States.

But those increasingly tempted to reach for a gun at the mention of the word bioethics should pause before doing so. It may be unrealistic to try to imbue the average laboratory scientist with the foresight needed to predict the full social impact of his or her discoveries, let alone their moral and ethical implications. But this does not absolve the researcher of all responsibility for thinking about the potential impact of those discoveries. It merely suggests that the link between discovery and application is a complex process in which both personal and collective judgements have important roles to play.

In this context, the important question to raise about the new steps being proposed to the European Commission is not whether they are either desirable or necessary — the global response to the reports of the cloning of Dolly the sheep should be sufficient to dispel any doubts on

both scores — but how they can be made to work most effectively. Here some important guidelines need to be followed.

First, it is essential that ethical assessments are based on sound science. One implication is that the scientific community has a clear responsibility to monitor the work of ethics committees at all levels, aware of the danger that, in highly charged debates, there can be a tendency for bad science to drive out good, sometimes exaggerating dangers for inappropriate reasons.

Second, the process must be sufficiently transparent to inspire the confidence of both the public and the research community. There is a need to overcome the culture of secrecy that, while perhaps protecting political reputations, can also undermine sound decision-making by limiting the scope of public scrutiny.

Third, there must be clear mechanisms to distinguish questions of ethics from those of safety. Although there are clear areas of overlap, there are also significant dangers of confusing the two in deciding how they should be handled.

Finally, an active effort must be made to eschew an excessively formalistic approach to ethical questions, remembering that one man's set of clear principles can rapidly become another man's moral and intellectual strait-jacket. A pragmatic approach has its own dangers; but these are usually easier to rectify than one that is cast in stone.

If these principles are ignored, there is a danger that what started as a well-intentioned effort to bridge the gap between the need to respect scientific freedom and legitimate public concern about the impact of some new scientific discoveries will have backfired. Excessive bureaucratic requirements that produce few concrete results will merely become one more impediment to Europe's scientific creativity.

But if these principles are followed, Europe will have demonstrated its ability to develop a sound approach to a key aspect of science policy appropriate to the political and social demands of the new millennium. Both science and the community that supports it will be better off. □

Preprints and *Nature*

Exposure of preprints on servers does not preempt their submission to this journal.

In some scientific disciplines it is common for draft papers to be circulated to colleagues before publication — a practice that has become formalized with the advent of electronic preprint servers pioneered at the Los Alamos National Laboratory. Superstring theorists and others working at high energies were in the vanguard, and it may be relevant to note that these communities are, within the spectrum of disciplines, unusual in their approach to peer review — for example, many high-energy physics experimental papers are refereed extensively within the large teams of collaborating authors before being circulated.

In other areas, where the communities are larger and the variability in quality and sheer volume of preprinted material somewhat daunting, preprint servers are active but appear to be considered less useful for those reasons. And in some areas of research, life is too fraught with competition for preprints to be advisable. Nevertheless, preprint servers are, for some, part of the means of routine scientific

discourse. They seem unlikely to displace the need for the selection and broader propagation of results provided by most journals.

There is no conflict between preprint circulation and submission to *Nature* — we only request that submitting authors inform us where and when a preprint has been placed on a server. But we do actively discourage prior exposure of results in the public media, given our conviction that the breadth and quality of coverage is strengthened by our embargo policy (for a statement of embargo policy and sanctions, see "author information" at www.nature.com). But *Nature* is more than just an archive. So we reserve the right not to publish papers whose results have already been disseminated and discussed beyond the ranks of specialists. And, as with conferences, our policy is that scientists wishing to publish in *Nature* are required not to encourage or cooperate with media coverage before publication, while communicating freely with other researchers in whatever way seems appropriate. □



US offers possible compromise at start of climate treaty talks

[KYOTO & LONDON] The first day of the United Nations conference on climate change, which opened in Kyoto, Japan, this week, saw a significant policy shift by the United States that could pave the way for an eventual agreement on targets to reduce emissions of the manmade greenhouse gases that can cause global warming.

US officials announced on Monday (1 December) that they are willing to consider an agreement that contains different greenhouse gas reduction targets for different countries, rather than a flat-rate target for all developed countries.

Until now, most developed countries have supported a flat-rate target even though sharp differences remain on what that target should be. Australia, Canada and New Zealand are proposing some of the weakest targets. Most developing states support a 15 per cent reduction in emissions from 1990 levels by 2010, as proposed by the European Union (EU). The United States, meanwhile, has proposed merely bringing emissions back to 1990 levels before 2012.

But Melinda Kimble, a state department official and a member of the US negotiating team at Kyoto, said on Monday that the United States is now willing to consider what is known as 'differentiation', "particularly if it would help resolve our remaining differences on the appropriate target level".

Further details are likely to emerge before US Vice-President Al Gore attends the conference next week. Gore confirmed his attendance only on Monday — but also said that the United States would be prepared to walk away from what it considered a bad treaty.

The 10-day conference is being attended by 10,000 participants comprising country representatives, nongovernmental organiza-



First words: Kyoto governor Teichi Aramaki addresses the opening session of the climate meeting.

tions and the media. This week's talks between government officials are designed to lay the ground for the arrival of ministers next Monday (8 December) for three days of negotiations. The deadline for agreement is midnight on 10 December.

As the meeting opened, significant obstacles to an agreement remained. These included whether countries should be allowed to trade in emissions credits, the question of developing country participation, and even the definition of what constitutes 'anthropogenic' emissions of greenhouse gases.

Europe and developing countries remain opposed to US proposals to let countries achieve their targets via a greenhouse gas trading regime. Developing countries want the current pilot phase to continue, and, although Europe is not opposed in principle to greenhouse trading, it wants the United States to raise its overall reduction target first.

But US support for differentiation does provide a way of resolving another contentious issue: its insistence that voluntary

targets be adopted by certain large developing states. Developing countries are exempt from reducing their greenhouse emissions at this stage under the terms of the UN climate convention, and want things to stay that way.

The United States has made clear that it will not sign an agreement at Kyoto unless larger developing countries such as India and China take steps to address their own emissions. According to Kimble, however: "Targets appropriate for developing countries need not take the same form as those under consideration for Annex 1 [developed] countries. They could take the form of emissions growth targets."

The United States used the opening day of the conference to step up its war of words with Europe. Kimble said: "The EU has not explained why it should not take on a more ambitious target in light of its unique economic advantage." This refers to the fact that EU member states such as Germany and the United Kingdom have achieved low emissions as the result of the closure of economically inefficient industries.

Europe is opposed to most of the US conditions for an agreement to limit emissions. Because of this it has enjoyed support from developing countries and plaudits from environmentalists. This has made US officials uncomfortable, and they used the first day of the conference to expose inconsistencies in the EU's proposal to limit emissions.

US support for differentiated targets puts the EU in an embarrassing position. Member states are opposed to differentiated targets. But their flat-rate target is the result of differentiated responsibilities between the 15 states. Some will reduce their emissions. Others, such as France, will stabilize. Poorer countries, such as Portugal and Greece, will be allowed to increase emissions to a certain level.

Asako Saegusa & Ehsan Masood

Transgenic patents a step closer in Europe

[MUNICH] The 'internal market' ministers of the member states of the European Union last week approved the latest draft of the European Commission's directive on the legal protection of biotechnological inventions, which allows the patenting of human genes as well as transgenic plants and animals.

The controversial directive has already been approved by the European Parliament in its first reading last July, following revisions to an earlier draft rejected two years ago (see *Nature* 388, 314; 1997). A text formally approved by the Council of Ministers is likely to be submitted to the

parliament early next year. Assuming — as most commentators do — a reasonably trouble-free ride through its second parliamentary reading, the directive is likely to be adopted by the end of next summer.

The council's approval has dismayed environmentalist and other anti-patenting pressure groups, which continue to oppose the directive on moral grounds. But it has lifted spirits in the European Patent Office, where work on patent applications for biotechnological inventions has halted because of uncertainty over whether its own rules allow patents on life.

Alison Abbott

Germany tightens grip on misconduct...

[MUNICH] Germany's Max Planck Society (MPS), which runs 73 basic research institutes, has approved a new set of internal regulations on the way in which cases of suspected scientific misconduct should be handled. And its senate has agreed that an educational programme in scientific ethics should be set up for young researchers.

According to the MPS president, Hubert Markl, this programme, to be set up through its Wissenschaftliche Rat (scientific council), reflects a belief that preventing misconduct is more fundamental than catching the relatively few scientists who go off the rails. Young scientists must be instilled with "a wide and deepened consciousness of the importance of being responsible members of the scientific community," he says.

Scientific ethics courses are likely to include such issues as correct ways of keeping scientific notebooks, criteria for deciding the authorship of a paper, and criteria for acknowledging technical contributions to papers. Issues relating to the social impact of research would also be discussed.

Parallel ideas about how to address scientific misconduct are emerging from the Deutsche Forschungsgemeinschaft (DFG), the major grant-giving organization for universities, which earlier this year found itself unprepared to deal with the most spectacular case of systematic scientific fraud to come to light in Europe.

In this case, two medical researchers, Friedhelm Herrmann of the University of Ulm and his former colleague Marion Brach of the University of Lübeck, have been accused of fabricating data in more than 40

publications (see *Nature* **387**, 750 & **389**, 105; 1997). Brach has admitted fraud, and says that Herrmann put her under pressure to cheat. Herrmann blames Brach, and claims he was unaware that papers he co-authored included data from experiments that were never carried out.

The case shocked Germany's scientific establishment, partly because of the extent of the deceit and partly because it took so long to come to light. Young researchers in Herrmann and Brach's laboratory had been aware for some time that publications from the laboratory included false data, but were afraid to make a complaint for fear of jeopardizing their careers.

The DFG, which had awarded substantial grants to the two researchers, then set up an international committee to study the issue of scientific misconduct, and will draw up a report based on its recommendations.

Smaller cases of fraud that previously came to light in Germany have been handled in an *ad hoc* way, and most members of the scientific community appeared content with this arrangement. One exception was Albin Eser, director of the Max Planck Institute for International Criminal Law in Freiburg. Until the Herrmann and Brach affair, Eser had been a relatively lone voice calling for standard procedures for such cases.

The recommendations of a committee set

up by the MPS in 1996 and headed by Eser have now been enthusiastically adopted by a newly sensitized Max Planck Society. The new procedures are designed to ensure that cases of scientific misconduct are dealt with rapidly, and that both whistleblowers and those innocently accused are protected.

Eser's committee has defined about 15 types of scientific misconduct, ranging from deliberate data manipulation and infringement of intellectual property rights to compromising the research of colleagues.

According to the new rules, when suspicion of misconduct is raised in a Max Planck research institute, the director should carry out an immediate informal inquiry within the institute, protecting the name of the whistleblower, and informing in confidence the MPS vice-president responsible for the research area. Anyone accused of misconduct will have two weeks to respond, and will be told within a further two weeks whether a formal investigation will be launched.

Formal investigations will be carried out by a new standing committee, whose chairman, elected by the senate, will have no connection with the MPS or its institutes. The committee will include the relevant MPS vice-president and three members of the society's existing arbitration committee.

The investigations committee will decide if misconduct has occurred and may make recommendations about sanctions. Alternatives would include a simple warning, a demand for return of grant money, dismissal or—in extreme cases—calling in the public prosecutor. The MPS president will decide which sanction to apply.

The report of the new DFG committee is likely to focus mainly on ways of ensuring that good scientific practice is followed. And, like the MPS, it is expected to emphasize the importance of attempting to solve problems at local level.

The DFG has advised universities to appoint independent counsellors to whom young scientists can turn when they suspect malpractice. The DFG president, Wolfgang Frühwald, has suggested that the DFG appoint an ombudsman to act as final arbitrator of unresolved cases.

Frühwald also suggested that scientific societies and universities draw up their own codes of practices. The German Physical Society has responded by setting up a committee to draw up a 'code of honour'.

Frühwald is concerned that political overreaction to the Herrmann and Brach case could lead to strong central controls on research. He is worried that such a move could stifle scientific creativity. But Eser believes that a national committee should eventually be set up "to streamline things" for all research institutions.

Alison Abbott



Frühwald: 'fraud ombudsman needed'.

FOTOSTUDIO QUERBACH

...while US students own up to cheating

[LONDON] Despite a steadily increasing emphasis on the need for scientists to act ethically (see above), cheating remains widespread among students at US universities, according to a survey of 4,000 students at 31 institutions.

The survey found that incidents of serious malpractice have increased significantly over the past three decades and, although highest among students on vocational courses such as business studies and engineering, they are also significant in the natural sciences.

The survey report, by Donald McCabe, professor of

management at Rutgers University in New Jersey, appears in the current issue of the journal *Science and Engineering Ethics* (**4**, 433–445; 1997). Based on the experience of university departments, McCabe concludes that strict penalties are a more effective deterrent than exhortations to behave morally.

Cheating is more common at universities without an 'honour code' – a binding code of conduct for students, with penalties for violation. More than half of science students at universities with no honour code admitted falsifying data in laboratory experiments.

More than two-thirds of all students polled said they had cheated in some way. Seventy-three per cent of science students from universities without an honour code admitted "serious cheating". The figure for those from universities with a code was 49 per cent. "Serious cheating" includes copying from someone during an examination, and using crib notes.

Cheating at honour-code universities was acknowledged by 73 per cent of students on business courses, 56 per cent of engineering students and 53 per cent of social science students.

Ehsan Masood

NASA looks to extended balloon flights

[WASHINGTON] The US National Aeronautics and Space Agency (NASA) hopes within the next few years to begin scientific balloon flights of up to 100 days duration as a way of conducting near-space research at a fraction of current launch costs.

The so-called Ultra-Long Duration Balloons would carry payloads of 2,000 kg or more, including about a tonne of scientific instruments. They would typically circle the globe at altitudes of 40 km, high enough for astronomical observations as well as studies of the stratosphere.

NASA already launches 20 to 30 scientific balloons each year, including a few long-duration flights. But balloons can at present stay aloft for only a couple of weeks.

The new, much larger 'superpressure' balloons would be made of advanced plastics such as Mylar or polyethylene, with a lightweight lamination to add strength. The lamination techniques were pioneered in sails for racing yachts.

Several tests of the superpressure balloons are scheduled in the next two years, and NASA plans a demonstration flight, with a scientific payload, in the year 2000. Six candidate experiments were selected last April, one of which will be chosen this winter for the flight. If the demonstration is successful, regular 100-day flights could begin in 2001.

Such flights would be "almost like a satel-



Longer life? Conventional balloons can only stay up for a couple of weeks before losing pressure.

lite" for some kinds of research, says Jack Tueller, balloon-project scientist at NASA's Goddard Space Flight Center in Maryland. High-altitude balloons are ideal platforms for infrared and γ-ray astronomy, and for studies of the cosmic microwave background.

And balloon experimenters usually get their scientific payload back, says Tueller. Their instruments can also have a wider variety of shapes and sizes, as they do not have to fit within the launch shroud of a rocket. Launch costs are another advantage of scientific balloons. A typical mission can launch for under \$1 million, says Tueller, compared with up to \$18 million using a Pegasus rocket.

NASA is working on designs for a 'ballooncraft' with standardized subsystems to provide power, communications and

other essentials to scientific instruments on 100-day flights, which would further reduce development costs.

The technical obstacles facing superpressure balloons are thought to be fairly straightforward, but the programme still faces political barriers. Tueller says the 100-day flights would probably launch from the Southern Hemisphere at first, largely for political reasons.

Circumnavigating the globe in mid-northern latitudes would require overflight permission from many countries, including Libya and Iraq. Missions flying at higher northern latitudes would pass over Russia, which has not yet agreed to allow such flights.

Safety issues will also have to be resolved before the balloon programme gets under way. Balloon payloads have been known to fall to the ground, and Tueller concedes that scientists would face a greater risk of losing their instruments during a 100-day flight than on shorter missions.

NASA intends to begin almost immediately to take advantage of this cheap way of reaching the edge of space. The agency will soon be looking for low-cost space missions for its University Explorer programme, and will for the first time include the option of flying instruments on long-duration (10-20-day) balloon flights as well as on satellites or the space shuttle.

Tony Reichhardt

Female scientists wanted – apply to UK research councils

[LONDON] Britain's six research councils are to keep a close eye on the relative success of men and women applying for grants and fellowships. The move coincides with the revelation that, although success rates are similar, there is clear evidence that a higher proportion of male than female academics apply for research awards.

Margaret Beckett, the President of the Board of Trade and as such, the cabinet minister responsible for science, said last week that she has asked the director-general of research councils, Sir John Cadogan, to report annually on the success rates of women in obtaining grants and fellowships.

Beckett holds a degree in metallurgy and worked briefly as a researcher before entering politics. Speaking at the annual meeting of the pressure group Save British Science, she said she was determined to encourage more women to enjoy a scientific career. "We run the risk of wasting the talents of half the population," she warned.

But she also said that it is necessary to ensure that women work in a fair and unbiased environment, adding that there had been "some disquiet" over a recent study of the competition for fellowships at the

Swedish Medical Research Council which revealed bias in favour of male candidates (see the Commentary by Wennerås and Wold, *Nature* 387, 341-343; 1997).

According to Beckett, a survey of success rates in British research councils carried out after the publication of the Swedish data did not reveal a similar pattern, the rates between the two sexes being "very similar".

Last year, for example, 24 per cent of male grant applicants to the Biotechnology and Biological Sciences Research Council, and 26 per cent to the Medical Research Council (MRC), were successful; equivalent rates for female candidates were 19 and 29 per cent.

Similar conclusions are revealed in a survey, *Women and Peer Review*, published this week by the Wellcome Trust's Unit for Policy Research in Science and Medicine. Also prompted by the Swedish study, the Wellcome survey concludes that there is "no evidence of sex discrimination" in the award of the trust's project or programme grants, or its senior research fellowships.

But both the Wellcome survey and a parallel survey by the MRC reveal that women do not apply for awards in the proportions that would be expected from the

number of female academics working in British universities (see Correspondence, page 438).

In the case of Wellcome, five times as many men as women applied to the trust, even though there were only marginally more male academic biomedical staff than female in the academic year 1995-96 (indeed, in the under-25 age group, there were 50 per cent more female biomedical academics than male). Similar statistics emerge from the MRC survey; only 21.3 per cent of applicants for programme grants were women.

The surveys' authors offer no explanations. But they recommend that research funding bodies should, as part of their efforts to ensure equal opportunities, work together to investigate the factors that may prevent women from applying for grants in the expected numbers.

Officials at the Department of Trade and Industry point out that efforts to redress the sex balance in research include the Dorothy Hodgkin fellowships, introduced by the Royal Society in 1995. They also argue that the new concordat for contract research should enable more women to remain longer in the academic community.

David Dickson

Governments and environmentalists split over forests

[LONDON] Sharp differences about the application of the United Nations (UN) biodiversity convention to forestry policy erupted between Latin American governments and environmentalist groups at a recent meeting of scientific experts from member countries of the convention.

At issue was whether the biodiversity convention should directly address the protection of biodiversity in forests, or be limited to providing advice to the separate UN Intergovernmental Panel on Forests, which regulates the commercial timber industry.

Environmental groups support direct intervention, as most biological diversity within forests lies in 'non-tree components', and the convention is more sympathetic than the panel to forestry conservation. But countries with large timber logging industries, particularly in Latin America, want the panel to take the lead role.

The declaration from the expert meeting, which took place in October, said that the biodiversity convention should be "consistent with" the panel's proposals for action. But one environmentalist at the meeting insists that "biodiversity should be given primary importance".

Environmentalists were also angry at being excluded from observing the drafting session. The chairman, Gabor Nechay from Hungary, took the unprecedented step of inviting environmentalist groups to help draft recommendations for governments. This contravenes UN rules, which say that only sovereign states can take part in drafting official documents.

The environmentalists' involvement was opposed by delegates from Argentina, Brazil, Colombia and Venezuela, who then succeeded in having the environmentalists excluded from the sessions altogether.

A UK member of an environmentalist group admits that the chairman overstepped his authority in allowing nongovernmental organizations (NGOs) to participate in drafting recommendations.

But he says the Latin American group had no right to exclude environmentalists from observing. NGOs can normally attend 'closed' sessions if participating governments agree. Such agreement depends on the sensitivity of the issue being discussed.

The Latin American countries' decision to exclude environmentalists is, however, unlikely to set a precedent. The new chairman of all the experts' working groups, Abdul Hamid Zakri from Malaysia, is known as a supporter of NGO involvement in the UN process.

Ehsan Masood

Canadian inquiry calls for 'safety first' blood agency

[MONTREAL] A controversial four-year inquiry into the infection of thousands of Canadians with HIV and hepatitis C through blood products has called for a new national blood agency, run by an independent authority with a mandate to make safety a high priority.

One of the government's first responses has been to set up a Blood Safety Council, headed by a haematologist, to include scientists and representatives of consumer groups. Its tasks will include monitoring the implementation of the 50 recommendations of the commission that carried out the inquiry, headed by Mr Justice Horace Krever.

Among these recommendations is that the chief medical officer of the new blood agency, to be known as Canadian Blood Services, should report directly to its board. Medical scientists had previously complained that their input into decision-making was subordinate to that of administrators (see *Nature* 384, 602; 1996).

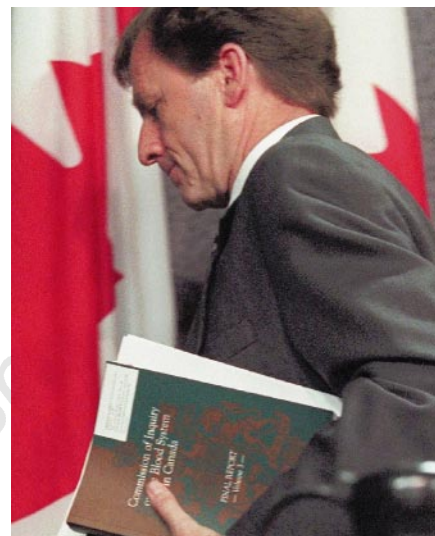
Provincial and federal governments had already begun to set up the new agency, which is expected to be in place by next September. Allan Rock, the federal health minister, said last week that Krever's more detailed recommendations would be taken into account in its design.

In making this move, the government appears to be back-tracking on earlier efforts to undermine the inquiry's conclusions. It had previously tried through court action to prevent Krever from blaming blood products for the infection of thousands of Canadians with HIV and hepatitis C (see *Nature* 379, 479; 1996).

The Canadian Blood Committee, one of the bodies charged with overseeing the blood supply, denied Krever access to potentially incriminating documents by destroying them in 1989. The government then set up its own task force to reform the blood system before Krever's report was presented.

Krever won an appeal to the Supreme Court that freed him to blame individuals but warned him to use language that would avoid litigation. Individuals are named in the report, with descriptions of actions Krever calls wrong or harmful. The Royal Canadian Mounted Police has announced a review to see if criminal charges are warranted.

Krever's report shows that the widely reported estimates of transfusion-related cases of AIDS (1,200) and hepatitis C (12,000) were far too low. It says that 85 per cent of hepatitis-C infections due to transfusions between 1986 and 1990 and 133 AIDS cases could have been avoided had available tests been used.



Rock: applying lessons of infected blood scandal following publication of inquiry report.

Krever suggests that the new service should continue to include provincial and territorial ministers, but not federal government, as its function as the safety regulator places it in a conflict of interest. The Red Cross will no longer be involved.

The separatist government of Quebec will not be a part of the new service, but instead will set up its own. But, according to Rock, Quebec's service will still be regulated for safety by the federal government.

Krever recommended that blood and blood products be paid for directly by hospitals, taking decision-making about needs away from politicians. He called for plans for a C\$300 million (US\$210 million) blood fractionation plant in Nova Scotia to be scrapped, but recommended that an amount equal to 10 per cent of the annual operating budget be allocated to research and development. The blood service should have its own research facilities, but should also collaborate with other organizations.

Krever also proposed a no-fault insurance system to compensate victims of contaminated blood. Since the 1980s 1,100 individuals infected with HIV in this way have received on average C\$30,000 a year, but those infected with hepatitis C have received nothing.

Following Krever's report, the government and the Red Cross have apologized to the victims of the blood tragedy. But the head of the Canadian Haemophilia Society, Durhane Wong-Rieger, expressed disappointment: "While the report is quite strong overall, we are disappointed that the level of responsibility and blame stops at the lower levels," she said.

David Spurgeon

European grants to face ethics scrutiny

[PARIS] European life scientists seeking research grants from the European Union (EU)'s next five-year Framework programme (FP5), due to start at the end of next year, may soon face an additional hurdle — the requirement that research proposals receive a stamp of 'ethical approval' from the European Commission.

The recommendation that the ethical implications of all such proposals be systematically evaluated is likely to form part of an 'opinion' on the commission's proposals for FP5 due to be released next week by the commission's Group of Advisers on the Ethics of Biotechnology.

The opinion is expected to justify such a move on the grounds that research is not 'neutral', and that its potential social and moral impact should be considered from the outset. It is expected to state formally that scientific freedom should not be given priority over risks to safety or human rights.

This argument drew a broad consensus at a meeting in Brussels last week of all the national ethics bodies of the EU member states, convened by the commission to provide input into the advisory group's opinion. "It is obvious that there is a tension between the freedom of science and the social responsibility of scientists, and that scientific liberty is not absolute," says group member Anne McLaren, a prominent reproductive biologist and former foreign secretary of Britain's Royal Society.

Although it is unclear how the suggested evaluation will operate in practice, the commission's proposals for FP5 already state for the first time that research supported should "comply with fundamental ethical principles". It explicitly bans funding for research on human germline gene therapy and human cloning, and states that animals used in experiments must where possible be replaced by *in vitro* or other alternative techniques.

One conclusion from last week's meeting was that the commission should not make ethical judgements on issues such as human embryo research where no consensus exists among member states. But it was also agreed that all proposals should be screened to comply with certain principles, such as the need to weigh up the benefits of animal research against the suffering caused.

The opinion is also likely to state that even if a research practice is forbidden in one country, that will not be sufficient to forbid it at European level. More controversially, the group is likely to recommend that research proposals potentially giving rise to broad ethical concerns, such as the search for genes related to homosexuality, should receive special in-depth evaluation. At the same time, the group recognizes that there are limits to the responsibility of researchers.

One recommendation that may prove embarrassing for the commission is the group's conclusion that fundamental research itself has an 'ethical value'. The opinion is likely to state that the EU has a moral responsibility to shift its funding away from applied research to fundamental research, on the grounds that the latter is being neglected in member states because of the emphasis on wealth creation. "The EU should be supporting research that might not be funded otherwise," says one group member.

More broadly, the commission's decision to bring together all national ethics committees underlines the growing political importance of bioethics in the EU. Jacques Santer, the commission president, who himself attended the meeting, said that a "balance" has to be struck between the economic prospects of new technology and the "human and social dimension".

Indeed, Santer said he intended to extend the group's remit to include all new technologies. The group itself is likely to recommend the incorporation of all the national ethics committees within a European ethics network under its umbrella. It is also likely to recommend closer links between the group and the European Parliament, whose members

have often criticized it as a 'puppet' of the commission.

According to Noelle Lenoir, chairwoman of the group, ethical evaluation at the European level is likely to be based more on the Anglo-Saxon 'pragmatic' approach of weighing up the pros and cons of particular cases, rather than the French approach of drafting legislation based on abstract principles such as the need to respect 'human dignity'.

Although the EU's founding treaties give it no formal competence in bioethics, the existence of directives on issues such as genetically modified organisms and gene patents makes litigation in such areas inevitable, according to Jean-Pierre Puissochet, a judge at the European Court of Justice.

The increased emphasis on bioethics within the EU is considered as inevitable and welcome by many observers. Some are concerned, however, that it will place unnecessary burdens on scientists and raises the risk of subjective assessment of research proposals on their political correctness. Projects "could be censored on grounds of public acceptability" warns one commission official, adding that after the BSE crisis, "everyone [at the commission] is trying to cover their backside".

Declan Butler

Seven-year watch on transgenic maize

[PARIS] The European Commission is to tighten a directive first issued in 1991 covering the deliberate release of genetically modified organisms into the environment. Revisions include the compulsory monitoring of products such as transgenic maize for harmful effects for seven years after their approval.

Under the proposed measures, full details of which are to be announced this week, the commission's scientific advisory committees will have to be consulted about all applications for the release of genetically modified organisms. Another provision allows the commission, for the first time, to refer dossiers to its Group of Advisers on the Ethics of Biotechnology (see above).

Ritt Bjerregaard, the environment commissioner, says that the new measures demonstrate that the



Bjerregaard: aware of public concern and debate.

commission "pays careful attention to public concern and public debate". The commission has already agreed that all products containing genetically modified organisms should be labelled, and is due to release details of labelling procedures this month.

The proposals, which need the approval of the ministers of the 15 member

states of the European Union, are aimed at breaking a deadlock over genetically modified crops. Although last year the union approved modified maize produced by Switzerland's Novartis, imports of the maize have been banned by Austria, Luxembourg and Italy on the grounds that they may pose health risks — a position challenged by the commission and France.

In a related development, the French government last week lifted a ban on the growing of genetically modified maize. But it said that further studies were needed on the production and sale of other crops, such as rapeseed and sugar beet.

Louis Le Pensec, the agriculture minister, said that the ban introduced by the previous government "was not coherent to authorize imports of genetically modified maize while prohibiting cultivation".

D.B.

Yeltsin voices concern over 'brain-drain'

[MOSCOW] The Russian president, Boris Yeltsin, has expressed concern that the continued exodus of scientists — and the resulting decline in the strength and prestige of Russian science — has become a threat to national security.

Presiding over a meeting of the Security Council in Moscow last week, Yeltsin pointed out that about 50 of the country's 100 best known scientists had left Russia.

"The 'brain-drain' still continues, although not as intensively as earlier," he said. "This is leading to dangerous consequences, both economic and political, since our intellectual and spiritual resources are being exhausted."

Yeltsin's concerns are based on the fact that, although a growing number of students and postgraduates are being temporarily sent abroad by the government to study at foreign universities, many postgraduates and researchers at Russian universities are choosing to leave the country permanently.

Yeltsin is reported to have started to discuss the brain-drain problem spontaneously — it was not on the agenda of the Security Council meeting. In an address lasting almost two hours, he also expressed concern that Russia's education system does not produce enough scientists. He warned members of the council that Russia may lose its technological independence as a result.

Yeltsin said he was dissatisfied with the efforts of federal ministries and other government bodies to reform science. He demanded that all scientific organizations immediately be paid all the money owed to them by the state, even within the present restricted budget.

A report prepared for Yeltsin by the Science Statistics Research Centre of the Ministry of Science and Technologies says that Russia is continuing to lose its strategic scientific potential at an increasing rate.

Since the beginning of the decade, 2,000 individuals engaged in science and education have left the country each year, and

overall Russian science has lost about 40 per cent of its top level researchers.

Unesco estimates the cost to Russia of this brain-drain as US\$30 billion. But the country no longer has the ability to restore such intellectual losses, as its universities receive only 5 per cent of government funds allotted for science.

This, in turn, means that public interest in science is low. A survey by the statistics centre showed that only 2 per cent of scientists expect young researchers to join their laboratories. And half of those questioned had not heard about recent scientific achievements such as cloning.

Commenting on Yeltsin's remarks, Vladimir Aleksandrov, deputy rector of Moscow State University, admitted that 10 per cent of its scientists are now abroad.

According to Vladimir Kryuchkov, former chief of the KGB (State Security Committee), "unofficial sources and unpublished information" to which he had access indicated that 200,000 scientists left the former Soviet Union between 1988 and 1991.

Vitaly Goldansky, a counsellor to the Russian Academy of Sciences, challenged this figure, suggesting that there were only "a few tens of thousands" of such people. But he admitted that "it is not the figure that matters, but the fact that they were the most able and talented scientists". **Carl Levitin**



US air base in Japan increases monitoring of local dioxin threat

[TOKYO] Officials at a US naval air base in Atsugi, Japan, plan to install an infrared sensor to monitor gases from a nearby waste incinerator which they claim is discharging toxic pollutants such as dioxin and tetrachloroethylene.

The problem is "the number one issue of concern for US forces in Japan," says an official at the US Naval Air Facility (NAF) base, which has more than 2,500 residents.

Although some local authorities have taken action against various sources of dioxin — for example, by ordering the closure of school incinerators — national regulations have yet to be implemented. Officials have refused to take action in the Atsugi case, claiming that the exhaust gases are within permitted limits.

The move by US officials to increase their own monitoring capabilities comes after Jinkampo, a privately owned waste incineration facility 250 metres from the NAF base, was granted a five-year extension of its waste disposal licence, despite objections from residents of the base.

US military officials have been analysing the emissions since 1988, after complaints from residents. Recent tests have allegedly

identified 12 pollutants emitted from Jinkampo. Officials concluded that the health risks were unacceptably high by US standards. The officials say that twice as many individuals suffer from respiratory diseases such as asthma at Atsugi as at other bases in Japan.

But appeals to the local prefectural government in 1989, asking for action to stem the emissions, and a report submitted to the Environment Agency in 1995 listing toxic substances detected in air samples, have been ignored by Japanese officials. They argue that emissions are within Japanese air pollution standards.

Officials at the base now intend to use Fourier transform infrared spectroscopy, which can monitor and analyse emissions from chimneys whatever the wind and atmospheric conditions, to improve their analyses of air samples. Data collected by the previous method, which used ambient air data, were not recognized by the Japanese government.

According to Brian Murphy, an environment specialist at NAF Atsugi, the main difficulty is that the waste incinerator is privately owned, and that Japan does not

have an "environment protection agency" to establish and enforce pollution laws.

Japan depends heavily on incinerators for waste disposal, burning 73 per cent of its waste, compared to only 16 per cent in the United States. In the United States, stringent air quality standards make it difficult for incinerators to comply with pollution laws.

The Japanese government has tried in the past to seek alternative methods of waste disposal. But opposition from industry has blocked implementation, partly because limited land means that methods other than incineration are expensive and complex. One result is that Japan regulates only ten hazardous air pollutants, compared with 195 in the United States.

The government has only recently started to introduce measures to regulate dioxin levels, following appeals from the public, including lawsuits against incineration plants for causing health hazards. Pollution caused by dioxin, a by-product of waste incineration suspected of causing cancer, is a major problem in Japan, where 450 million tonnes of waste are disposed of each year. **Asako Saegusa**

UK plans for coping with plutonium waste fail to impress

[LONDON] More than 100 members of Britain's Parliament last week signed a motion criticizing the government for failing to develop a policy for managing plutonium removed from dismantled nuclear weapons. The MPs also opposed the decision to convert part of the plutonium stockpile into mixed oxide fuel (MOX) for nuclear power reactors.

The MPs want to halt plans to build a £500-million (US\$800-million) MOX recycling plant until a review of policy has been completed. Their public statement coincides with the publication of the report of a two-year study, coordinated by the Citizens' Nuclear Information Center in Tokyo and the World Information Service on Energy in Paris, which concludes that recycling MOX could lead to plutonium proliferation.

MOX is made by mixing oxides of uranium with small quantities (5–10 per cent) of oxides from weapons-grade plutonium. Some countries, including Britain, France, Germany and Japan, are considering the manufacture of MOX as a way of reducing plutonium stockpiles. They also believe it to be a useful alternative to fossil fuels.

The authors of the international MOX assessment study argue that such a strategy could be counterproductive. They contend that it will keep plutonium in circulation for longer than necessary, increasing the risk of its misuse by terrorists. They also point out that the quantity of plutonium in the MOX mixture is too small to make a dent in the stockpiles. France has over 40 tonnes of plutonium stockpiled, and Britain over 50 tonnes.

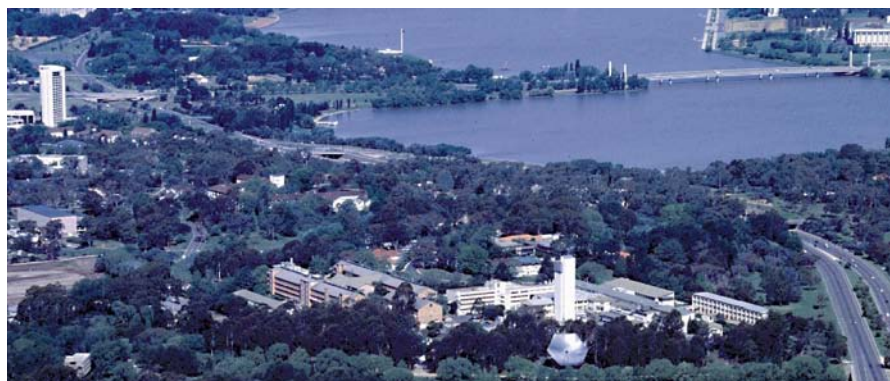
The study also raises questions about security at plutonium stores, claiming that site operators, particularly in Britain, lack the capability to trace material that is stolen or illegally diverted to weapons making.

Site operators aim to keep 'unaccounted-for material' to below 8 kg, the minimum quantity estimated by the International Atomic Energy Agency to be needed to make a nuclear device. But the MOX study claims that the Thermal Oxide Reprocessing Plant operated by British Nuclear Fuels at Sellafield in northwest England cannot detect missing plutonium unless the material weighs more than 220 kg.

The study claims that a minimum weekly inventory of plutonium stocks is needed to detect stolen or diverted material quickly "instead of every year as practised today".

But such claims are rejected by Peter Metcalfe of British Nuclear Fuels. "The plutonium is subject to very strict checks and safeguards by multinational teams of inspectors to ensure that it is safely handled and accounted for," he says.

Ehsan Masood



Troubled waters: the ANU campus, with the Institute of Advanced Studies in the centre foreground.

Staff resist 'streamlining' at top Australian institute

[CANBERRA] Tension is running high at the Australian National University (ANU) in Canberra, where efforts to absorb a 14 per cent cut in government funding over four years have included a move to change the management of its renowned Institute of Advanced Studies (IAS).

The past two months have seen both the closure of the university's Division of Archaeology and Natural History and the resignation of Sue Serjeantson, one of its two deputy vice-chancellors, who had also been director of the IAS since 1993.

The archaeology group had built an international reputation for its work on the pre-history of Australia's indigenous people (see for example, Roberts *et al.*, *Nature* 387, 696–699; 1997). The group's fate is seen as an example of the damage that can be done when the government cuts university funding without a clear policy.

Founded as a research-based university 51 years ago, ANU recently emerged from two major external reviews of the IAS with an acknowledgement of the quality of research in its seven schools, and with little change to its unusual funding arrangements or the federated structure of its teaching and research departments (known as 'The Faculties').

A unique feature of the IAS, which has enabled it to attract researchers of international standing, has been the arrangement under which it receives an annual block grant of \$A150 million (\$US105 million) from the government, two-thirds of the total grant to the ANU.

The resulting autonomy has allowed the IAS (1,590 staff out of 3,570 for the whole of ANU) to set its own priorities and carry out long-term studies unaffected by the short-term perspective forced upon researchers in other universities.

Although Serjeantson has not revealed the reasons for her resignation, which followed a review of ANU's senior management structure by Peter Karmel, a former ANU

vice-chancellor, it is widely believed to have been triggered by a perceived threat to the special characteristics of the institute. Karmel's panel proposed abolishing her post on the grounds that the university's administration should be more streamlined at a time of "uncertainty, competition, change and constrained funding".

Karmel's recommendation of a merger between the managements of the IAS and The Faculties sparked vehement protests by staff concerned at a lack of consultation. They also claimed that the university's reputation would suffer if the traditional 'collegiality' of decision-making was replaced by 'managerialism' and the privatization of some of its activities in the search for nongovernment funding.

These claims have been rejected by the vice-chancellor, Deane Terrell, who describes the management plan as "an effective way of integrating collegial academic responses to issues" which will "protect and enhance Australia's pre-eminent research groups in the IAS".

Under a structure revised following objections, the institute will continue to have a director. But although the post will be at the same salary as a (sole) deputy vice-chancellor, the latter will have direct control over resources and responsibility for research in both the IAS and The Faculties.

A vote last month at a meeting of the university's governing council was tied between academics seeking to defer the plan and administrators in favour of immediate implementation. The casting vote of the university's chancellor, Peter Baume, allowed Terrell to win the restructuring, but he gave the two academic boards a month's grace to comment on details.

Both boards describe the plan as "confusing" and want the council to restore responsibilities to the director of the IAS and the parallel post of chair of The Faculties board.

Peter Pockley

German reunification hits grant money for young scientists

[MUNICH] Young scientists in Germany are becoming increasingly disheartened by a sharp decrease in the likelihood of being awarded a research grant by the Deutsche Forschungsgemeinschaft (DFG), the main grant-giving agency for university research, as a direct consequence of the country's reunification. That is the conclusion of a report published by the DFG last week.

Since 1989, the DFG has found itself under pressure to provide the same level of support to east German university scientists as to their west German colleagues. As a result, despite an annual 5 per cent increase in budget since 1990, its overall level of support has fallen.

DFG was able to approve 80 per cent of grant applications in 1989, yet can manage only two-thirds today.

Pressure on the DFG coffers comes not just from having to absorb east German researchers but also from having to deal with more applications from west Germans, required to seek external research money because the public funding of research in Germany's universities has decreased in the post-reunification recession.

Since 1989, the value of applications for DFG grants has grown twice as fast as the organization's budget.

CNRS plan to encourage more long-term research

[PARIS] France's fundamental research agency, the 26,000-strong Centre National de la Recherche Scientifique (CNRS), is considering steps to reduce pressure on its scientists to 'publish or perish' in a bid to encourage them to explore more speculative long-term research avenues. Under plans being considered by the agency, research grants would be reviewed only every four years instead of every two as at present.

Many senior CNRS officials believe that the current system tends to encourage 'safe' research, as this is considered more likely to yield results within the two years. But it offsets one of the advantages of the French system, that the life tenure enjoyed by all publicly funded researchers should in principle give them an unrivalled opportunity to carry out long-term research.

Centre renamed to honour Abdus Salam

[LONDON] The International Centre for Theoretical Physics in Trieste, Italy, has been renamed the Abdus Salam International

Centre for Theoretical Physics in honour of its Nobel-prizewinning founder Abdus Salam, who died a year ago after a long illness (see *Nature* 384, 296 & 520; 1996).

Salam, who was born in Pakistan (then part of India), set up the centre in 1964 to help physicists from developing countries. It was funded by the International Atomic Energy Agency in Vienna, Unesco, and the Italian government, which now provides most of the funding. The centre has since expanded, and offers facilities for scientists from a range of disciplines.

India and United States strengthen health links

[NEW DELHI] India and the United States have signed a five-year agreement on collaborative research in the field of contraceptives and reproductive health. The two countries have also decided to extend by another five years the Indo-US Vaccine Action Programme, which was launched in 1987 amid sharp controversy (see *Nature* 328, 287; 1987).

Both the new initiatives and an earlier US proposal for an Indo-US Science and Technology Forum are being seen as attempts to create a favourable climate for a scheduled visit by US President Bill Clinton to New Delhi early in 1998. The forum, whose membership is being finalized, will allow scientists of the two countries to

continue to collaborate after the end of the year, when support from the US-India Fund, through which joint projects have been funded since 1987, comes to an end.

Euro parliament brings FP5 a step nearer

[MUNICH] After spending many long nights trying to rationalize nearly 700 proposed amendments to the European Commission's draft proposal on FP5, the European Union's fifth Framework programme of research (1998–2002), the parliamentary research committee last week managed to agree a structure for FP5, but not a price.

The committee proposes an increase in the number of thematic programmes suggested by the commission from three to four, and was due to meet on 4 December to try to decide on the financing. It is likely to propose a total budget slightly higher than the ECU16.3 billion (US\$18.3 billion) proposed by the commission.

Scott takes leading role in New Zealand science

[SYDNEY] Sir George Scott, a prominent neurophysiologist who retired last January as professor of medicine at the University of Auckland, has been elected president of the Royal Society of New Zealand. The society's

scope has recently been enlarged by an act of parliament from traditional sciences to include engineering, and information, mathematical, medical, social and technological sciences.

The change has increased the total membership of the society to 60 societies — with about 20,000 members — and 244 elected fellows. George Petersen, a biochemist from the University of Otago, Dunedin, has been elected president of the society's Academy Council.

Brookhaven lab's new managers named

[WASHINGTON] A partnership between the State University of New York at Stony Brook and the Battelle research corporation will be awarded the contract to run the troubled Brookhaven National Laboratory in New York state, the US Department of Energy has announced. A new management team headed by John Marburger, former president of Stony Brook, will take over in January.

Marburger will be joined by half-a-dozen senior managers from the Pacific Northwest National Laboratory in Washington state, which is run by Battelle. The 3,400 existing staff at Brookhaven will keep their jobs, Marburger says, although the responsibilities of some senior staff will change. The new team's first challenge will be to advise

Federico Peña, the energy secretary, on whether to try to reopen the High Flux Beam Reactor, which was shut down earlier this year after a leak of tritium.

Japan's delayed space missions go into orbit

[TOKYO] Japan's National Space Development Agency (NASDA) last week launched two missions on board the H-II rocket from Tanegashima Space Centre on an island off Japan's south coast (see below). These were the Tropical Rainfall Measuring Mission (TRMM), a joint project between the US space agency NASA and Japanese space institutes, and NASDA's Engineering Test Satellite VII (ETS-VII). The launch had been postponed for more than a week because of an anomaly in TRMM's data subsystem.

Both ETS-VII and TRMM went into their



Up and away: Japan's latest H-II launch.

prescribed orbits as planned, and TRMM deployed solar panels and has begun solar acquisitions. But an anomaly in the main solar paddle drive system means that the ETS-VII experiments will have to be suspended until the paddles have been fixed.

NASDA

No evidence of sexism in peer review

Sir — A recent study of peer-review scores for postdoctoral fellowships at the Swedish Medical Research Council demonstrated that women had to be 2.5-times more productive than their male colleagues to get the same peer-review rating in competitions for personal fellowships¹.

Using the same approach, we have audited recent applications to the Wellcome Trust and the UK Medical Research Council to examine whether there is discrimination in awarding practices^{2,3}.

The table presents the outcomes of grant applications for fellowship awards and project grants in the Wellcome Trust and MRC. There is no *prima facie* evidence that women are discriminated against in the awarding process — the award rates for both sexes are approximately the same.

Neither is there any evidence that women need a more impressive publication record than men to be successful in either organization's competitions. Bibliometric analysis demonstrated that there was no statistically significant (that is, $P \geq 0.05$) difference in the number of papers published by successful men and women. For example, female candidates awarded project grants by the trust published, on average, 11.2 papers in the five years preceding the application, compared to 13.8 papers for men.

As an indirect indicator of quality, the final column in the table shows the average journal impact factor for the journals in which the papers were published⁴. For the Wellcome Trust's project grants and the MRC's Career Development Awards there was no significant difference between men and women. However, in the Senior Research Fellowships in Basic Biomedical Science men reaching the interview stage published in higher impact journals than women ($P < 0.05$).

Another finding is that women do not apply for research funding in the numbers that might be expected from the numbers of women employed in academic positions in medical and biosciences departments in UK universities. For example, for project grants, where the sex of the applicant was known, only 19.6% (268/[1,097+268]) of applicants to the trust and 21.3% (167/[617+167]) of applicants to the MRC were women. Yet across all UK universities, 44% of academic staff (whether research oriented or otherwise) in medicine and biosciences are women^{2,5}.

In conclusion, this study has shown no evidence of discrimination against women in the assessment of applications for Wellcome Trust and MRC personal fellowships or grants. Nevertheless, it is notable that so few

The outcome of biomedical grant applications

	No of applications†	Success rate (%)	Av. No of publications‡ (over a 5-year period)	Expected 5-year impact§ (av. citations/paper)
Wellcome Trust				
<i>Project Grants (1996)¶</i>				
Men	1097	27.5	13.8	11.7
Women	268	26.9	11.2	13.3
Total	1387	27.5	12.5	12.5
<i>Senior Research Fellowships in Basic Biomedical Science (1994–95/96–97)¶¶</i>				
Men	220	5.5	14.3	29.3
Women	105	8.6	11.8	22.7*
Total	354	5.9	13.5	27.0
Medical Research Council				
<i>Project Grants (1996)¶</i>				
Men	617	26.4	–	–
Women	167	29.3	–	–
Total	850	25.6	–	–
<i>Career Development Awards (1993–96)</i>				
Men	169	16.6	5.0	22.3
Women	107	10.3	4.6	28.6
Total	276	14.1	4.9	24.1

* $P < 0.05$

† The sex of some applicants was not recorded.

‡ Papers retrieved from the Science Citation Index (SCI) and Social Science Citation Index (SSCI) were limited to articles, notes and reviews in accordance with normal bibliometric analysis of substantive research outputs.

§ Five-year citation records, taken as the average number of citations received by items published in 1990 and cited in journals processed for the SCI/SSCI in the years 1990–94.

¶ The bibliometric analysis of successful project grant applicants was restricted to a random sample of 25 men and 25 women.

¶¶ To overcome problems with small numbers, the bibliometric analysis of Senior Research Fellowships in Basic Biomedical Science is based on those invited to interview (39 men and 20 women).

female academic staff are applying for research funding, and it is important that the reasons for this are examined further.

Jonathan Grant

*The Wellcome Trust,
210 Euston Road,
London NW1 2BE, UK*

Simon Burden

*Gillian Breen
Medical Research Council,
20 Park Crescent,
London W1N 4AL, UK*

1. Wennerås, C. & Wold, A. *Nature* **387**, 341–343 (1997).
2. *PRISM Women and Peer-review: An audit of the Wellcome Trust's decision making process* (Wellcome Trust, December 1997).
3. <http://www.mrc.ac.uk>.
4. *Journal expected citations rates file* (Institute for Scientific Information, Philadelphia, 1990).
5. These data were commissioned from the Higher Education Statistics Agency (HESA). Biomedical science is defined as all full-time academic staff (researchers and other grades, lecturers, senior lecturers and readers, and professors, whether or not tenured) working in the following cost centres: clinical medicine, clinical dentistry, veterinary science, anatomy and physiology, nursing and paramedical studies, health and community studies, psychology and behavioural sciences, pharmacy, pharmacology and biosciences.

The hole truth

Sir — The assertion by Rolf Müller and eminent others (*Nature* **389**, 712; 1997) that the term 'ozone hole' was "coined in the mid-1980s" is in error by at least 50 years.

In his presidential address to the Royal Meteorological Society in 1934 Sydney Chapman (*Q. J. R. Meteorol. Soc.* **60**, 127–142; 1934) radically proposed the creation of ozone holes so that astronomers could perform better ultraviolet observations.

Chapman even documented a formal definition: "By 'making a hole in the ozone layer' I mean the removal of all, or most of, the ozone from the column of air resting on some chosen area."

Further, he suggested the possibility of inserting gas or fine powder into the stratosphere from aircraft, rockets and balloons and made the notable prediction that a catalytic "deozoniser" would be needed.

No doubt he would have taken great interest in the recent letter from M. N. Ross *et al.* (*Nature* **390**, 62; 1997), "Observation of stratospheric ozone depletion in rocket exhaust plumes", and recommended that ultraviolet astronomers should set up camp near a major rocket launching site.

Bob Wells

*Atmospheric, Oceanic and Planetary Physics,
Clarendon Laboratory,
Parks Road,
Oxford OX1 3PU, UK
email: wells@atm.ox.ac.uk*

Multi-ringed revelation

H. J. Melosh

The Chicxulub crater is the remnant of a meteorite impact that is blamed for killing off the dinosaurs. A new seismological survey shows that it is smaller than many scientists had thought, and extends deeper than anyone had suspected.

Planetary science conferences for several years have been disturbed by sometimes acrimonious exchanges between rival investigators of the Chicxulub impact structure. Created by a giant meteorite impact 65 million years ago, this huge scar in the rocks beneath Mexico is implicated in the death of the dinosaurs and other biological mayhem that ended the Cretaceous era. The arguments are not centred around the reality of the impact — that was settled to nearly everyone's satisfaction years ago — but on just how big the structure really is, and therefore how lethal the event was. Now, on page 472 of this issue¹, the rivals appear to have buried the hatchet and joined together behind a new and surprising interpretation of the structure, one that ties it more closely to some of the largest impact scars observed on the Moon and other planets.

The Earth is constantly bombarded by interplanetary debris. Within this fusillade of mostly small meteoritic fragments, a comet or asteroid more than 10 kilometres in diameter strikes the Earth roughly once every 100 million years, and may cause a mass extinction^{2,3}. Of the 150 or so known impact craters on Earth, only three were created by objects this large. Chicxulub is the most recent of the three: Sudbury in Ontario, Canada, is 1.85 billion years old⁴; and Vredefort, in South Africa, is 2.02 billion⁵. Both of these ancient craters are deeply eroded and their original structures are now obscure. But although Chicxulub is by far the youngest and best preserved of the large

craters on Earth, it is not particularly accessible. In fact, its very existence was not widely known outside the confidential files of various oil companies until 1991. The problem is that the crater does not outcrop anywhere at the surface.

Straddling the northern coastline of the Yucatán Peninsula, half below land and half below water, Chicxulub is buried by about a kilometre of post-Cretaceous sediments and must be studied by expensive drilling programmes or indirect geophysical methods. One of these is gravity⁶: an impact fractures the rock, lowering its density and creating circular gravity deficits around the crater; interior gravity highs may be caused by dense solidified impact melt or uplifted rings of basement rock. So it was natural that the first investigations of the structure of Chicxulub^{7,8} should make use of gravity anomaly measurements. But gravity data are not susceptible to unique interpretation, and the two main groups studying gravity anomalies at Chicxulub came to completely different conclusions about the structure.

They even disagreed on the size of the 'transient crater', the hole initially blasted out by the meteorite impact. As its name suggests, this initial cavity does not last long — in terrestrial craters larger than about three kilometres the transient crater is unstable, and in seconds or minutes collapses to produce a progression of new crater forms, which are determined mainly by the size of the transient crater.

The energy of the impact can be computed

ed⁹ from the transient crater diameter raised to the power of about 3.4, so it is important to obtain an accurate estimate of this dimension. The interpretations based on gravity anomalies alone gave transient crater diameters that ranged from about 90 km (ref. 7) to a whopping 170 ± 25 km (ref. 8), and thus impact energies differing by about a factor of ten.

The new work employs a much more powerful geophysical tool: deep seismic-reflection sounding. Developed by the oil industry, this technique produces images of structures deep underground using scattered seismic waves. The lead group, based at Imperial College, London, in collaboration with BIRPS (British Institutions Reflection Profiling Syndicate), imaged vertical slices through the structure along four lines that cross its offshore portion. Unfortunately, the centre of the structure is onshore, and the profiles could not delineate this interesting region. Nevertheless, these magnificent images seem to have resolved the controversy over the crater's true structure, and revealed some unexpected details.

The new data show clearly that the transient crater diameter was about 100 km and thus support the lower energy of impact (Fig. 1). The profiles show an inner ring of uplifted breccia (rock that has been broken up and stuck back together again) with a diameter of about 80 km. This appears to be a 'peak ring', an inner mountainous ring observed in large craters on every planet. Also, blocks of pre-existing rocks are seen that slid inwards and downwards, corresponding to the slumped and terraced rims of large extraterrestrial craters.

But the main surprise was that the prominent 170–195-km ring is not the crater's slumped rim, but instead a fault scarp (or more gradual slope, on one of the profiles) with a vertical offset of about 0.5 km. The inward-dipping fault plunges under the crater at 30–40° from the horizontal and visi-

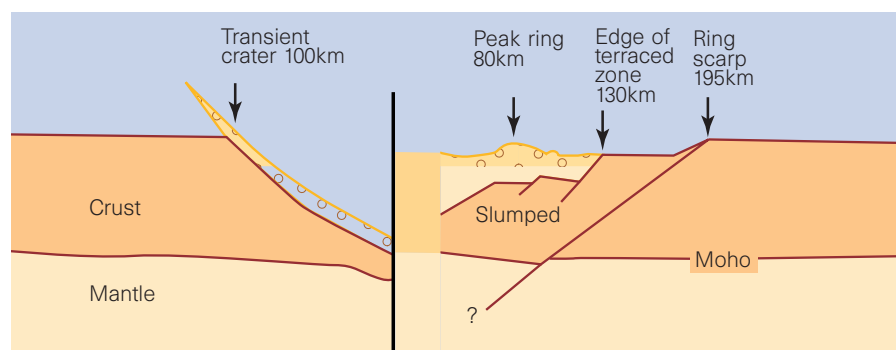


Figure 1 The Chicxulub structure, just before collapse of the transient crater, and after crater formation (fault displacements greatly exaggerated). A new seismic survey has fixed the size of the transient crater, and found that some faults extend all the way into the mantle — something that appears to contradict conventional models of the crust.

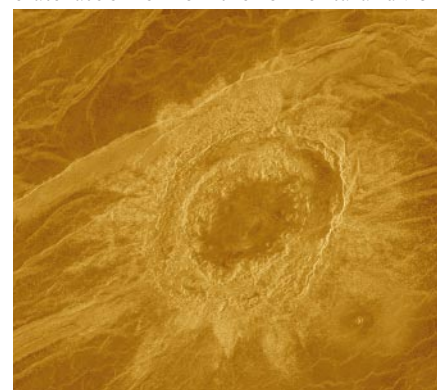


Figure 2 Venusian contusion — the 150-km-diameter multi-ringed impact basin Klenova, similar to many others on Venus and the Moon. New seismic data show that the Chicxulub crater is another such, the first to be discovered on Earth.

ROBERT STROM

bly offsets the Moho discontinuity (where the crust contacts the mantle) at a depth of about 35 km, after traversing the entire crust. Such inward-facing scarps are not new to planetary science. Craters on the Moon and Venus (Fig. 2) have such multiple rings. Strangely, however, even the largest craters on Mercury lack these faulted rings. They probably form when crustal blocks slide inwards over a weak layer¹⁰, and so the presence or absence of such rings may be telling us more about the target structure than about the impact¹¹ — implying, for example, that Mercury has a thick crust.

This is the first time such a ring has been identified in any terrestrial crater (although there are hints that the Sudbury structure might once have been surrounded by a deep-seated ring¹²). At slow plate-tectonic rates of deformation, the lower crust is a weak zone, sandwiched between the brittle upper crust and stiff upper mantle — but the Chicxulub faults cut the entire crust and extend an unknown depth into the upper mantle. So ring theorists must now find the ‘weak layer’ responsible, or an entirely new explanation of multi-ringed impact basins. For example, ring formation must have been very rapid (within minutes of the impact) because the fault scarp is buried in crater breccia, but does not offset it, so the faults must have formed before the breccia. The effective strength of layers beneath the crater at high strain rates might not be the same as for the low-strain-rate tectonic deformation more familiar to plate dynamicists. As in other areas of geodynamics, the nature of ‘strength’ depends

on many circumstances that are not fully understood.

The Chicxulub structure is smaller and more interesting than previous interpretations had suggested. It is the only proven multi-ring impact basin on Earth, and it shows that this type of structure is surprisingly deep-seated. Chicxulub is sure to lead to new revelations in the mechanics of impact crater collapse. But one question is settled: the impact produced a transient crater only 100 km in diameter. If that was enough to cause the Cretaceous/Tertiary extinction, it appears that impacts can be more devastating to Earth’s biosphere than we supposed. □

H. J. Melosh is in the Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA.

e-mail: jmelosh@lpl.arizona.edu

1. Morgan, J., Warner, M. & The Chicxulub Working Group *Nature* **390**, 472–476 (1997).
2. Morrison, D., Chapman, C. R. & Slovic, P. in *Hazards Due to Comets and Asteroids* (ed. Gehrels, T.) 59–91 (Univ. Arizona Press, Tucson, 1994).
3. Rampino, M. R. & Haggerty, B. M. in *Hazards Due to Comets and Asteroids* (ed. Gehrels, T.) 827–857 (Univ. Arizona Press, Tucson, 1994).
4. Krogh, T. E., Davis, D. W. & Corfu, F. in *The Geology and Ore Deposits of the Sudbury Structure* (eds Pye, E. G., Naldrett, A. J. & Giblin, P. E.) 431–446 (Ontario Geol. Surv., Ontario, 1984).
5. Moser, D. E. *Geology* **25**, 7–10 (1997).
6. Pilkington, M. & Grieve, R. A. F. *Rev. Geophys.* **30**, 161–181 (1992).
7. Hildebrand, A. R., Pilkington, M., Connors, M., Ortiz-Aleman, C. & Chavez, R. E. *Nature* **376**, 415–417 (1995).
8. Sharpton, V. L. *et al.* *Science* **261**, 1564–1567 (1993).
9. Melosh, H. J. *Impact Cratering: A Geologic Process* (Oxford Univ. Press, New York, 1989).
10. Melosh, H. J. & McKinnon, W. *Geophys. Res. Lett.* **5**, 985–988 (1978).
11. McKinnon, W. B. & Melosh, H. J. *Icarus* **44**, 454–471 (1980).
12. Spray, J. G. & Thompson, L. M. *Nature* **373**, 130–132 (1995).

carbon dioxide with increasing species number, but no further increase beyond an intermediate level of species richness. Just as with the grassland communities⁶, this probably reflects the importance of having the right mix of functional groups for average ecosystem performance. The new twist in McGrady-Steed and colleagues’ experimental results is that variability among replicates, and variability in time, decreased across the full range of species richnesses studied. Encouragingly, a re-analysis² of the previous studies^{4,5} gave a similar result.

Naeem and Li¹ started their communities with different numbers of species per functional group, chosen at random from the species pool. They observed an increasing variance in the biomass with decreasing species richness. This may be due to the larger number of species mixes that could be created with a smaller number of species per functional group, because the most diverse treatment included most of the species in the species pool. If so, species composition also matters — not just the composition of functional groups. (It would be valuable to repeat the experiment with a much larger species pool.) On the other hand, the larger number of species in the more diverse treatments may have allowed more fine-tuned compensatory growth, and thereby increased community-level predictability. In any case, the key conclusion of both studies is that the variance in ecosystem processes may continue to decline with increasing species richness, even if the corresponding average value would level off at an intermediate level of species richness.

Important and exciting as these results are, they do not reduce the value of similar studies on communities of plants and animals. Biodiversity in microbial communities obeys laws that sound peculiar to ecologists working with creatures visible to the naked eye. In one study, Fenchel and co-workers⁷ set out to find the total number of ciliated protozoan ‘species’ in two small sediment samples from an English lake and a Danish marine bay, covering an area of only 50 cm². They recorded all of the species that could be microscopically observed and, because rare or encysted species are likely to be missed in standard microscopic examinations, sub-samples were incubated under a range of environmental conditions. Thus, they created the missing ‘niches’ in which the species could reproduce and become measurable. The results were startling — initially, 68 species were identified, but eventually the minute sediment samples produced 247 species, comprising 75 per cent of the species known from the study localities and nearly 10 per cent of all known ciliate species in the world. And these are still minimum estimates, because the authors may not have created the conditions needed to ‘awaken’ all of the species in their samples.

Ecology

Be diverse, be predictable

Ilkka Hanski

One of the grand challenges of ecology is the prediction of the variability in the sizes of populations, in the species composition of communities, and in the function of ecosystems. How much this variability is affected by biodiversity — and by which kind of biodiversity — has remained especially contentious, and the political debate about biodiversity has attached extra weight to this.

Two papers^{1,2}, including one by Naeem and Li on page 507 of this issue¹, report inspiring results on how the predictability of microbial communities may depend on species richness. Taking advantage of the short generation times of algae (producers), bacteria (decomposers), and protists and small metazoans (consumers), the two groups constructed replicated microcosms (Fig. 1) with different levels of functional diversity and species richness. Naeem and Li show that an increase in the number of species per functional (trophic) group leads

to an increase in the predictability of biomass and density measurements. And McGrady-Steed and co-workers² found that temporal and among-replicate variation in the production of carbon dioxide (which is a measure of community function) decreases with increasing species richness. Both studies conclude that the presence of ‘redundant’ species in these experimental communities leads to more consistent ecosystem function — and this has obvious implications for the current controversy about the value of biodiversity.

Previous studies on grassland plant communities^{3,4} and an experimental multi-trophic-level community⁵ were thought to show how productivity increases with species richness. Subsequent analyses⁶ suggested that what matters most is the diversity of functional groups, with species number per functional group being less important. Similarly, McGrady-Steed *et al.* found a substantial initial increase in production of

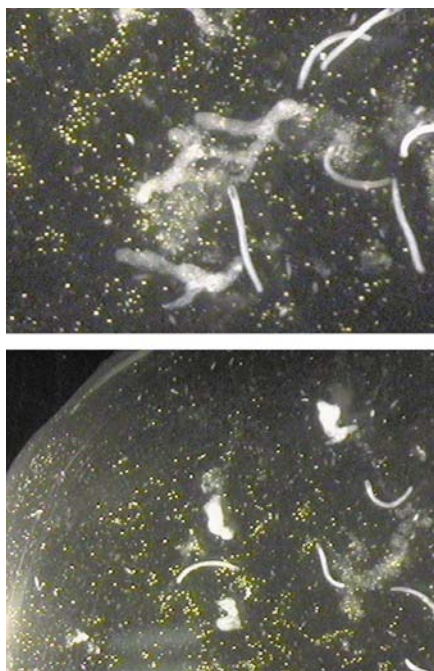


Figure 1 Community life on a small scale — microbial microcosms, similar to those used by Naeem and Li¹. Studies of such microcosms indicate that predictability of microbial communities depends on species richness. Naeem and Li constructed microbial microcosms containing autotrophs, omnivores and top predators. They found that, with increasing number of species per functional group, biomass and density measurements become more predictable.

Microbial communities therefore seem to have the peculiar property that “everything is (almost) everywhere”⁸. But most species are waiting for the turn in the environmental conditions that will allow them to flourish. This has led Finlay *et al.*⁹ to conclude that the concept of redundancy has little meaning for microbial communities — microbial diversity is never so impoverished that the microbial community as a whole would become a ‘limiting factor’ in the function of an ecosystem. This does not mean that experiments on microbial microcosms could not be used to explore general questions about community-level consequences of biodiversity.

Unfortunately, species of plants and animals cannot stay locally dormant and wait for their niches to reappear (with some important exceptions). When their niches disappear, plants and animals disappear. Extinctions will reduce the pool of ‘redundant’ species that could potentially buffer community and ecosystem function in the face of population dynamic variation. Extinctions will also prevent species from regaining their non-redundant roles if there is a more radical change in the environmental conditions.

The studies by Naeem and Li¹ and McGrady-Steed *et al.*² address the variability

of community and ecosystem processes, not variability in the abundance of individual species. From studies on grasslands, Tilman¹⁰ has suggested that although biodiversity may increase the stability of communities and ecosystems, it may also be associated with elevated variability of individual species, indicating substantial compensatory population growth in these species. This is also what seems to happen in microbial communities^{1,9}. In diverse natural communities, instability of individual species may reflect dynamical fragility of strongly connected complex systems¹¹. Or it could reflect the greater average degree of specialization of species in species-rich than in species-poor communities. This would make species-rich communities especially sensitive to environmental change.

The challenge of resolving issues in the diversity–stability puzzle has become less formidable with studies such as these, which

have begun to dissect it into smaller pieces. The microbial microcosms — which can be assembled from many species at several trophic levels and run for many generations — add an important tool to the ecologist’s arsenal. □

Ilkka Hanski is at the Department of Ecology and Systematics, University of Helsinki, PO Box 17, 00014 Helsinki, Finland.

e-mail: ilkka.hanski@helsinki.fi

1. Naeem, S. & Li, S. *Nature* **390**, 507–509 (1997).
2. McGrady-Steed, J., Harris, P. M. & Morin, P. J. *Nature* **390**, 162–165 (1997).
3. Tilman, D. & Downing, J. A. *Nature* **367**, 363–365 (1994).
4. Tilman, D., Wedin, D. & Knops, J. *Nature* **379**, 718–720 (1996).
5. Naeem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H. & Woodfin, R. M. *Nature* **368**, 734–737 (1994).
6. Tilman, D. *et al. Science* **277**, 1300–1302 (1997).
7. Fenchel, T., Esteban, G. F. & Finlay, B. J. *Oikos* **80**, 220–225 (1997).
8. Fenchel, T. *Oikos* **68**, 375–378 (1993).
9. Finlay, B. J., Maberly, S. C. & Cooper, J. I. *Oikos* **80**, 209–213 (1997).
10. Tilman, D. *Ecology* **77**, 350–363 (1996).
11. May, R. M. *Stability and Complexity of Model Ecosystems* (Princeton Univ. Press, 1973).

Solar System

Twice in a blue moon

Jane Luu

The inventory of the Solar System became a bit larger on 31 October, when a team of astronomers announced the discovery¹ of two new moons of Uranus, S/1997U1 and S/1997U2. To the relief of some astrophysicists, this makes the Uranian system more like those of the other giant planets.

The new satellites are ‘irregular’ satellites, in that their orbits are relatively far from the planet, with large eccentricities and inclinations compared with the ‘regular’ satellites.

They orbit Uranus at a mean distance of roughly 5.8 million kilometres (227 Uranian radii) — although the orbital parameters are less certain for U1 than for U2 — and are estimated to have diameters of 80 and 160 km, respectively². News of the discoveries was greeted with relief: until now, Uranus was the only giant planet without irregular satellites of its own (Jupiter currently holds the record with eight irregulars; Neptune has two and Saturn one). Scientists generally do not like deviations from patterns, and are pleased

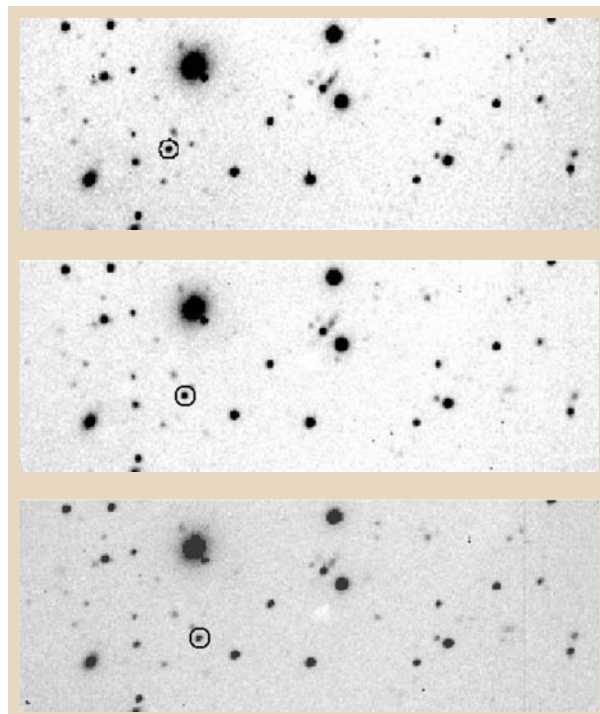


Figure 1 The discovery of U2. The circled dot is the newly discovered Uranian moon S/1997U2, distinguishing itself from the background stars by its movement. (Photos taken about an hour apart.)

that Uranus has now joined the fold, making irregular satellites a common trait of all giant planets. But however pleased astronomers may be, they cannot help but be reminded that the origin of irregular satellites eludes them.

It is generally accepted that, unlike the regular satellites, the irregular satellites did not form *in situ*. The regular satellites accreted within circumplanetary disks, but the irregular satellites must have been captured from elsewhere, most likely from orbits around the Sun. So how can a body initially in an unbound orbit (from the planet's perspective) become bound to the planet?

Capture theories fall mainly into two categories. In one, the satellites were captured by dissipative forces such as gas drag or tidal friction³. Such forces could also fracture a single captured satellite to form a cluster of satellites — Jupiter has two of these clusters. The second category of capture theory involves a collision between an asteroid and a satellite⁴: the collision shatters one or both bodies, creating fragments that are then captured into planetocentric orbits. Still other hypotheses rely on *ad hoc* assumptions about the initial mass and orbit of the planet in order to enable satellite capture⁵. But no single explanation has been able to account for the dissimilar sizes and configurations of the irregular satellites.

It is too early to tell how the new Uranian moons will improve our understanding of the origin of satellites. Perhaps, with better-determined orbits, the dynamical characteristics of the moons will rule out certain formation theories while adding credence to others. But the new moons are just two new pieces of a large puzzle. Recent years have seen a burst of new discoveries in the outer Solar System, most of them in the trans-Neptunian region⁶. This is now known to be the realm of the Kuiper Belt⁷, a large population of icy bodies that counts Pluto among its members, and probably the Neptunian satellite Triton as an ex-member. Instead of being mere empty space, the outer Solar System has turned out to be a rich and complex region, shaped by inward and outward migrations that have persisted since the earliest days of planet formation^{8–10}.

Perhaps some or all of the irregular satellites migrated inwards from the Kuiper Belt. This was probably the case with Triton, and further observations of the new Uranian moons and of Kuiper Belt objects might tell us whether it is more generally true.

So, although the two new Uranian moons are welcomed for making irregular satellites ubiquitous among the giant planets, the relief is short-lived: they are also a sobering reminder that even the contents of our own Solar System are not known. And if we are to understand planetary systems around other

stars, we should first understand what our own planetary system contains, and why it looks the way it does. □

Jane Luu is in the Department of Astronomy, Harvard University, 60 Garden Street, Cambridge, Massachusetts 02138, USA.
e-mail: luu@cfa.harvard.edu

1. Gladman, B. J., Nicholson, P. D. & Kavelaars, J. J. *IAU Circ. No.* 6764 (1997).

Structural biology

The enzyme at the end of the food chain

Richard Cammack

Whenever biomass is degraded, and oxygen runs out, the methanogenic bacteria appear and make methane in profuse quantities. For us it is a potential fuel. For the microbes, it is the ultimate waste product. The structure of the enzyme that produces it, methyl-coenzyme M reductase (MCR), has been worked out by Ermiler *et al.*, and published in *Science*¹. The structure of the nickel centre and its interaction with substrates reveal the mechanism of one of the few organometallic reactions in biochemistry.

Most microbes, other than photosynthetic ones, obtain their energy for growth by converting metabolites into compounds of lower free energy, which they then excrete. Often other microbes can break these compounds down further. When there is no oxygen present, the end of this chain is the realm of the methanogens², their monopoly on methane production apparently stemming from their unique cell structure and metabolism³. The ultimate scavengers, they take carbon compounds such as acetate and CO₂, and wring the last quantity of energy from them by reducing them to methane.

Some methane is released into the atmosphere (concentrations are currently about 1.8 parts per million by volume), and it is a significant greenhouse gas. However, we are generally unaware of the huge amounts of methane that are made in the biosphere (estimated to be over 10⁹ tonnes per year; ref. 4), because as soon as it comes into contact with oxygen, most of it is reoxidized to CO₂ by methanotrophic bacteria using another unique enzyme, methane monooxygenase⁵. Only in special circumstances are the attentions of the methanotrophs avoided — for example in the digestive tracts of herbivores, including termites, and in anaerobic soils such as rich paddies. Annual methane release through these routes is respectively estimated at 10⁸ and 2 × 10⁸ tonnes.

The conversion of CO₂ to methane is an integrated sequence of seven steps, involving unusual cofactors, unstable enzymes and

2. Marsden, B. G. & Williams, G. V. *IAU Circ. No.* 6765 (1997).
3. Pollack, J. B., Burns, J. A. & Tauber, M. E. *Icarus* **37**, 587–611 (1979).
4. Colombo, G. & Franklin, F. A. *Icarus* **15**, 186–189 (1971).
5. Saha, P. & Tremaine, S. *Icarus* **106**, 549–562 (1993).
6. Jewitt, D. C., Luu, J. X. & Chen, J. *Astron. J.* **112**, 1225–1238 (1996).
7. Kuiper, G. in *Astrophysics* (ed. Hynek, J. A.) 357–424 (McGraw-Hill, New York, 1951).
8. Ip, W.-H. *Icarus* **80**, 167–178 (1989).
9. Torbett, M. *Astron. J.* **98**, 1477–1482 (1989).
10. Levison, H. & Duncan, M. *Icarus* **127**, 13–32 (1997).

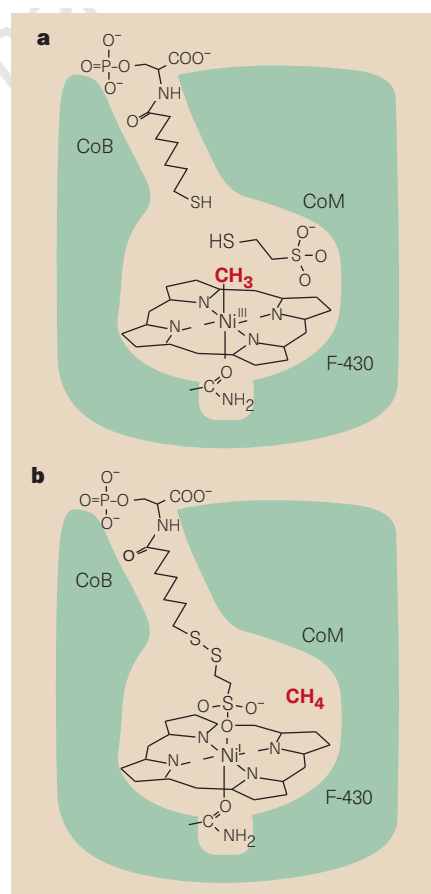


Figure 1 Location of coenzymes in the active site of methyl-coenzyme M reductase. The methyl group and methane are absent from the crystal structures, but their possible location is indicated. a, The methyl group is transferred from coenzyme M (CoM) to nickel. b, Methane is released and a disulfide is formed between CoM and coenzyme B (CoB). Release of methane and the mixed disulfide completes the reaction cycle.

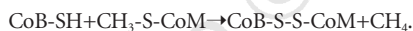
complex interactions with transmembrane ion gradients. Elucidation of the pathway^{1,6} is a notable achievement of modern biochemistry.

The studies concerned have revealed how methanogens produce methane, and why. Almost all processes of biochemical energet-

ics involve an oxidation–reduction reaction (even fermentations can be considered as internal oxidation–reduction reactions). For the methanogens, living in extremely reducing conditions, reductants such as hydrogen (H_2) are present in abundance; the problem is to find, or generate, a sufficiently strong oxidizing agent. The purpose of their elaborate metabolic pathway is to generate a disulphide between two coenzymes (B and M), which, although not a good oxidant by most standards, is oxidizing enough. The re-reduction of the disulphide by H_2 , in a subsequent enzyme reaction in the cell membrane, leads to energy production in the form of a proton gradient and eventually ATP. The methane is essentially a by-product of the production of the disulphide, catalysed by MCR.

From its first isolation by Wolfe and his group⁷, there were indications that MCR is an unusual enzyme. It is a large molecule (relative molecular mass around 300,000). It has to be isolated in the absence of oxygen, and even then it soon loses activity. Ermiler *et al.*¹ have determined, to very high resolution (1.4 Å, refined), the structures of two different inactive forms of the enzyme. They have further demonstrated the relationships of these states to the active enzyme⁸. From this it is possible to deduce what the active enzyme looks like, and how it might catalyse its reaction.

The reaction catalysed by MCR involves three unique organic cofactors⁹. Two of these are organic cosubstrates, coenzyme M (CoM, 2-thioethanesulphonate) and coenzyme B (CoB, 7-thioheptanoyl-threonine-O-phosphate):



MCR is one of the few enzymes to contain nickel¹⁰, in the form of factor F-430, a tetrapyrrole complex. In MCR the proximal ligand to the nickel, uniquely, is the carbonyl oxygen of a glutamine. The other ligand position is believed to be the site at which the methyl group of methyl-CoM is converted to methane. In bacterial metabolism, methyl groups are conventionally handled by vitamin B_{12} -type coenzymes, which have cobalt in a tetrapyrrole complex, and which do not release methane. The reason for using nickel instead of cobalt in MCR is that it binds the methyl group less tightly, so that methane can be released by protolysis.

The structure shows the small cofactor CoM sitting inside the enzyme in a chamber with the nickel cofactor F-430 at the base. In one of the crystalline forms of MCR, the sulphide of CoM is coordinated to the nickel. Presumably, in the active enzyme, methyl-CoM, in a similar configuration, could donate a methyl group to the nickel (Fig. 1a). In the other crystalline form, the CoM molecule has flipped round, with its sulphonate group facing the nickel, and forming a disulphide with CoB (Fig. 1b).

A narrow channel, 30 Å long, leads from the chamber to the surface. The channel is particularly inert and hydrophobic. CoB has a side chain of seven carbon atoms, which plugs the channel while the reaction takes place. Contrary to previous expectations, the -SH group of CoB cannot get closer than 8.7 Å to the nickel. If it could make contact, it would coordinate to the nickel and might be difficult to remove again. During the proposed reaction cycle, the nickel undergoes changes in oxidation state, as observed by electron paramagnetic resonance¹¹, and then releases methane and oxidizes the thiol groups of CoB and CoM. The most plausible mechanism for such a reaction is a free-radical process involving long-distance electron transfer, through the protein, from the nickel to the thiolate. Water is probably excluded from the active site during this process, a feature of several enzymes with free-radical mechanisms¹.

Ermiler and colleagues' high-resolution electron-density map¹ suggests that some unprecedentedly modified amino acids lie near the active site. A thioglycine, having C=S instead of C=O in the peptide bond, is located in the active site, near to the -SH group of CoB, which might assist in the oxidation–reduction reaction. In addition, four different amino acids (a histidine, arginine, glutamine and cysteine) are methylated; these have not been seen in proteins before, and there is no known mechanism

for their insertion into the polypeptide sequence. An intriguing possibility is that they might be formed by methyl radicals produced during the free-radical reaction of the enzyme.

The methanogens belong to the Archaea, one of the ancient divisions of organisms². It is interesting that their methane-generating system involves nickel and free radicals, which are relatively rare in enzymes of the present day. So the new structure also gives a glimpse into what may have been one of the earliest forms of metabolism. □

Richard Cammack is in the Centre for the Study of Metals in Biology and Medicine, Division of Life Sciences, King's College, London W8 7AH, UK.

e-mail: richard.cammack@kcl.ac.uk

1. Ermiler, U., Grabarse, W., Shima, S., Goubeaud, M. & Thauer, R. K. *Science* **278**, 1457–1462 (1997).
2. Woese, C. R. & Wolfe, R. S. *Nature* **19**, 273–274 (1986).
3. Ferry, J. G. (ed.) *Methanogenesis — Ecology, Physiology, Biochemistry and Genetics* (Chapman & Hall, New York, 1993).
4. Hogan, K. B., Hoffman, J. S. & Thompson, A. M. *Nature* **354**, 181–182 (1991).
5. Rosenzweig, A. C., Frederick, C. A., Lippard, S. J. & Nordlund, P. *Nature* **366**, 537–543 (1993).
6. Thauer, R. K. *Antonie van Leeuwenhoek Int. J. Gen. Mol. Biol.* **71**, 21–32 (1997).
7. Ellefson, W. L. & Wolfe, R. S. *J. Biol. Chem.* **256**, 4259–4262 (1981).
8. Goubeaud, M., Schreiner, G. & Thauer, R. K. *Eur. J. Biochem.* **243**, 110–114 (1997).
9. Dimarco, A. A., Bobik, T. A. & Wolfe, R. S. *Annu. Rev. Biochem.* **59**, 355–394 (1990).
10. Hausinger, R. P. *Biochemistry of Nickel* (Plenum, New York, 1993).
11. Rospert, S., Voges, M., Berkessel, A., Albracht, S. P. J. & Thauer, R. K. *Eur. J. Biochem.* **210**, 101–107 (1992).

Geophysics

Going halves over Hudson Bay

Jerry X. Mitrovica

Regional variations in the gravitational pull at the Earth's surface can provide glimpses into the Earth's internal structure. One such anomaly (a negative one) is centred over Hudson Bay in North America (see cover), and has a maximum amplitude of roughly 1 part in 20,000 of the surface gravitational acceleration — which as anomalies go is quite sizeable. A definitive explanation for this anomaly has proved elusive. But on page 500 of this issue¹ Simons and Hager describe a fruitful new approach to the question, which not only provides the most direct answer yet, but should also be invaluable in understanding dynamic processes shaping the Earth.

There are at least two different explanations for the large Canadian anomaly, and they are not mutually exclusive. One stems from the remarkable correlation between its spatial form and the geometry of the Laurentide ice sheet during the last glacial maximum. This massive ice sheet covered northern North America for tens of thousands of years until its retreat some 10,000 years ago. The ice load would have caused compression

and outward viscous flow of the Earth beneath it — an event from which the Earth is still recovering in a process known as post-glacial rebound, and which is plausibly reflected in the gravity anomaly.

Qualitative appearances can, however, be deceptive: a series of studies have suggested that the anomaly may instead primarily be a manifestation of mantle convective flow below the North American craton. According to this view, North America lies above an anomalously cold, dense region of downwelling convective flow that acts to depress the surface topography. (The low temperature is one interpretation for anomalously high seismic velocities beneath the continent.) The free-air gravity anomaly results from the competing effects of the dynamically supported depression and the mass excess associated with the high-density mantle material.

Uncertainty about the two explanations has led geodynamicists to an impasse. Should gravity measurements in previously glaciated regions be used merely as checks on models of the Earth? Or can they provide important constraints in formulating such models?

Simons and Hager¹ have carried out an innovative analysis of the global free-air gravity field which explores the geographical variation in the spectral content of the field. The approach provides fresh insights and a powerful, gravity-based datum for inferring Earth structure. Specifically, as regards Hudson Bay, it confirms proposals that the post-glacial rebound of Canada is consistent with a significant increase of material strength (or viscosity) with depth in the mantle, and shows that 50% of the peak anomaly from this region is due to incomplete post-glacial rebound.

Associating the entire observed peak free-air gravity anomaly over Canada with post-glacial disequilibrium is highly problematic. It has long been recognized^{2,3} that the relatively large gravity anomaly is inconsistent with the short relaxation time (some 2,000–3,000 years) inferred from the post-glacial uplift record from the same region. Simply put, the uplift record indicates that the post-glacial adjustment is nearly complete, whereas the gravity anomaly suggests that a significant level of disequilibrium remains. To reconcile this apparent inconsistency, Walcott² proposed an adjustment defined by two relaxation times; however, Cathles³ argued convincingly that Earth models characterized by a double relaxation time would not fit other glaciation-related data sets in the region. Cathles was, to my knowledge, the first to suggest that a significant portion of the observed anomaly may derive from an entirely different source — namely, mantle convection.

The idea of a double relaxation time was revived by Peltier and Wu⁴ in the early 1980s. These authors argued that the buoyancy associated with the deflection of the density discontinuity at 670 km depth, the boundary between the upper and lower mantle, would be characterized by long decay times and would contribute significantly to the peak anomaly over Hudson Bay. Subsequent analysis⁵ showed that the free-air anomaly had been incorrectly calculated and that the contribution from the adjustment of the 670-km boundary would be very small. It now appears inescapable that the amplitude of the Canadian anomaly cannot be explained by remnant post-glacial disequilibrium alone^{6,7}. Indeed, direct predictions of the mantle convection signal, using methods established to predict long-wavelength variations in the geoid⁶, indicate that convective flow may support over 75% of the observed anomaly.

How can this uncertainty be resolved? One approach is to simply ignore the regional gravity data from such areas and test preferred models by their ability to account for the observed peak gravity signature. Forte and I inverted data associated both with mantle convection and with post-glacial rebound, and found that a single viscosity profile, characterized by a significant increase with depth, is able to reconcile the two⁸. Our

Palaeoclimatology

Climate's carbonate cypher

For the best part of 50 years, palaeoceanographers have been exhuming marine sediments and meticulously decoding the isotopic signals locked into the ancient carbonate shells of buried animals, such as the microscopic foraminifer, *Orbulina universa*, pictured here. The fruits of this cryptographic endeavour are detailed pictures of past climates and ocean conditions stretching back hundreds of millions of years. But have the right codes been used? Not always, according to new experiments on *O. universa* and another foraminifer, *Globigerina bulloides*, reported elsewhere in this issue (H. J. Spero *et al. Nature* 390, 497–500; 1997).

Pioneers in this field worked out that the oxygen-isotope ratio, $^{18}\text{O}/^{16}\text{O}$, of shell carbonate depends not only on the ratio in the surrounding sea water, but also on the temperature at which the shell formed. A shell's carbon-isotope ratio, $^{13}\text{C}/^{12}\text{C}$, seems to reflect that of the ambient sea water, overprinted with the effects of physiological processes, such as respiration and the photosynthesis of symbiotic organisms.

It is no secret that other processes affect



isotope fractionation, but their identity and extent of influence have proved largely elusive. Spero *et al.* show that the concentration of dissolved carbonate ions in sea water has a marked effect on the fractionation, and consequently argue that present understanding of some past climate changes will need to be revised. For example, the glacial–interglacial shift in shell $^{13}\text{C}/^{12}\text{C}$ — thought to be due to a large transfer of terrestrial carbon to the ocean — could simply reflect known changes in the surface ocean's carbonate ion concentration; and oxygen-isotope-based estimates of tropical sea-surface temperature during the last glacial period might need to be lowered.

Philip Newton

calculations indicated that the predicted peak free-air gravity anomaly fitted the observed anomaly, with one-third of the signal arising from post-glacial disequilibrium.

Simons and Hager¹ outline a second approach. They use a so-called spatio-spectral localization method⁹, which involves spectral decomposition of data spatially windowed around a specific geographical site. Repetition of the procedure for sites around the globe provides a measure of the geographical variation of this spectral content. The method gives an incredibly rich representation of the gravity field (see Fig. 1 on page 501), in which tell-tale fingerprints of various geophysical signatures (such as variations in crustal thickness, and mantle convection and incomplete rebound) are clearly evident. Simons and Hager find that, over Hudson Bay, the amplitude of the spatio-spectral representation of the gravity field is uniquely localized in space and in length scale. They also show that near-surface seismic tomographic images do not have this feature in the vicinity of Hudson Bay, which provides convincing evidence that the post-glacial rebound contribution to the gravity field in this region cannot be small.

But how large is this signal? The answer is an outcome of a second aspect of Simons and Hager's analysis, in which they define a global transfer function, $\bar{F}$, determined by spatially averaging local correlations between the

observed gravity field and any arbitrary spatial distribution. If this distribution reflects the spatial geometry of post-glacial rebound, then the transfer function is a much less 'contaminated' measure of the rebound signal than either a peak gravity anomaly or a traditional spectral decomposition. Indeed, the global transfer function derived by Simons and Hager (see their Fig. 2b on page 502) is an important new data set in studies of post-glacial rebound and one that can be used directly to infer Earth structure.

As a first step in this direction, the authors compute synthetic transfer functions using a set of previously published viscosity models as well as a new model ('SH') which provides a reasonable fit to their estimate of $\bar{F}$. The SH model shares many of the main features of the viscosity profile inferred by Forte and me⁸, including a stiff asthenosphere, an anomalously weak transition zone and a high-viscosity deep mantle (see Fig. 2a, page 502). This agreement is remarkable because the data sets used in the two studies are essentially independent, and it adds strong support to a growing consensus in the geodynamics community that the mantle is characterized by a large increase in viscosity with depth. It is on the basis of the SH model that Simons and Hager argue that 50% of the peak free-air gravity anomaly over Hudson Bay is due to incomplete glacial rebound.

Overall, the authors have resurrected the

use of regional gravity fields in studies of post-glacial adjustment and have demonstrated the great usefulness of spatio-spectral localization in inferring Earth structure. Potential applications extend well beyond the problem of glacial isostatic adjustment, and include research into variations in crustal thickness and the signature of mantle convection. One immediate issue involves the non-rebound signal in the peak gravity anomaly over Hudson Bay. The high seismic velocities below the North American craton arise from some combination of chemical and thermal heterogeneity. The gravity anomalies supported by these two sources of heterogeneity are distinct, and the free-air gravity field may prove to be an important datum for discriminating between them¹⁰. □

Jerry X. Mitrovica is in the Department of Physics, University of Toronto, 60 St George Street, Toronto, Canada M5S 1A7.

e-mail: jxm@physics.utoronto.ca

1. Simons, M. & Hager, B. H. *Nature* **390**, 500–504 (1997).
2. Walcott, R. J. *Can. J. Earth Sci.* **7**, 716–727 (1970).
3. Cathles, L. M. *The Viscosity of the Earth's Mantle* (Princeton Univ. Press, 1975).
4. Peltier, W. R. & Wu, P. *Geophys. Res. Lett.* **9**, 731–734 (1982).
5. Mitrovica, J. X. & Peltier, W. R. *J. Geophys. Res.* **94**, 13651–13671 (1989).
6. Peltier, W. R., Forte, A. M., Mitrovica, J. X. & Dziewonski, A. M. *Geophys. Res. Lett.* **19**, 1555–1558 (1992).
7. James, T. S. *Geophys. Res. Lett.* **19**, 861–865 (1992).
8. Forte, A. M. & Mitrovica, J. X. *Geophys. Res. Lett.* **23**, 1147–1150 (1996).
9. Simons, M., Solomon, S. C. & Hager, B. H. *Geophys. J. Int.* **131**, 24–44 (1997).
10. Forte, A. M., Dziewonski, A. M. & O'Connell, R. J. *Science* **268**, 386–388 (1995).

Reproductive biology

Do males rely on female hormones?

Richard M. Sharpe

It is true that 'hormones maketh the man'. Without the effects of androgens (the prime example of which is testosterone), the adult male is largely demasculinized in appearance, behaviour and drive. In the womb, androgens produced by the fetal testis divert development from the (default) female pathway along the male pathway. Without this hormonal 'intervention', genetic males would develop as phenotypic, but infertile, females. In contrast, adult females make large amounts of oestrogens which literally shape the female form during puberty. The concept that oestrogens mean female whereas androgens equal male is therefore true but, predictably, it is not this simple.

Both sexes make androgens and oestrogens. A minor role for oestrogens in the male — in regulation of gonadotrophin secretion from the pituitary gland — is accepted. But oestrogens may exert more widespread effects in the male and, remarkably, they may even be essential for fertility. This stems from the discovery that transgenic male mice lacking the α form of the oestrogen receptor (ER- α gene knockout, or ERKO, mice) are infertile¹. Now, a report by Hess *et al.*² on page 509 of this issue provides the first mechanistic clue as to why this may be so.

When sperm are released from their 'nurse' (Sertoli) cells in the testis, they are transported in fluid secreted by the Sertoli cells to a collecting area, the rete testis. From here, the dilute suspension of sperm enter the thin-walled efferent ducts, the epithelia of which express enormous amounts of oestrogen receptor^{3,4}. The sperm then pass into the epididymis where they mature and are prepared for storage. This involves the resorption of about 90 per cent of the fluid that surrounds the sperm, mainly in the efferent ducts. But Hess *et al.* have found that

fluid resorption is abnormal in the ERKO mice — rather than resorbing fluid, their efferent ducts secrete it, if anything.

The Sertoli cells continuously secrete fluid under androgen control. Because this fluid is not resorbed in the ERKO mice, it accumulates at the site of production — just as happens when you get a blocked waste pipe and run the tap. This build-up progressively impairs sperm production because of the increased fluid pressure within the testis. It is also likely that any sperm that manage to survive in ERKO mice may not mature properly in the epididymis, because the abnormal amounts of fluid will effectively dilute factors secreted within the epididymis which act on the sperm.

To identify how oestrogens exert their effect in the efferent ducts, the next step will be to unravel the biochemical details of fluid resorption, which is presumably centred on

ion and water channels. Several such channels are expressed in the efferent ducts — Na⁺/K⁺-ATPase channels⁵, chloride channels⁶ and aquaporin-1 (ref. 7). But which (if any) of these are regulated by oestrogens is not yet known. We should also expect some twists to this story, such as evidence that the oestrogen that controls fluid resorption may come from the sperm themselves⁸ and that, under some circumstances, oestrogens may inhibit rather than promote fluid resorption⁹.

Although it is easy to see how the chain of events following impairment of fluid resorption could lead to infertility in the ERKO male, this may not be the complete explanation. The sperm produced by these ERKO males are abnormal in their fertilizing ability¹, and this defect may be a consequence of the changes described above. Or it could be a signal that oestrogen action is important in normal spermatogenesis. Indeed, if the distribution of oestrogen receptors is anything to go by, we can expect many more actions of oestrogens to be identified in male reproduction. Such receptors — either ER- α or the more recently discovered ER- β — are expressed, alone or together, throughout the male reproductive tract: in Sertoli cells (which regulate sperm production)¹⁰; Leydig cells (the site of androgen production)³; the efferent ducts^{3,4}; the epididymis; and the accessory sex organs (which make the ejaculate)^{10,11}. Moreover, expression occurs not just in the adult, but in the fetus/neonate as well. Add to this expression of oestrogen receptors in the pituitary gland, brain, bone, skin, gut, heart vasculature and many other tissues, and we are faced with the likelihood that oestrogens have pervasive functions in the male.

Already, functions in the male that were previously ascribed to androgens are now known to be the result of oestrogen action¹². One example is closure of the epiphyses, which stops growth of the long bones in late

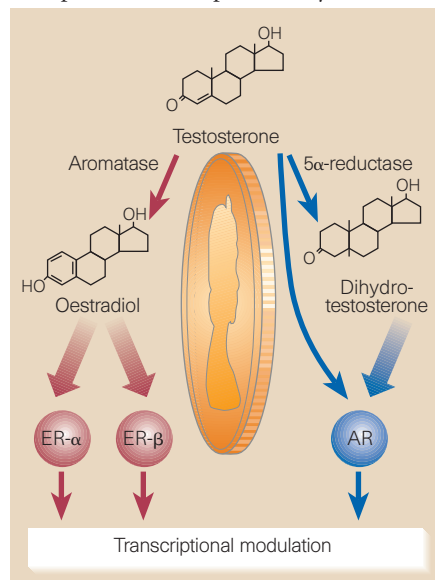


Figure 1 Oestrogens and androgens — two sides of the same coin. Hess *et al.*² have shown that transgenic mice lacking the α form of the oestrogen receptor have impaired sperm production due to a defect in resorption of fluid in the efferent ducts. These results indicate a previously unknown function for oestrogens in the male reproductive tract. More functions are likely to be discovered, because the androgen receptor (AR) and α and β forms of the oestrogen receptor (ER- α and ER- β) are co-expressed in many cell types in the male reproductive tract. This co-expression is not altogether surprising, however, as most androgens can be converted into oestrogens.

puberty and thus determines final height. Three human males have been discovered who either cannot make oestrogens or who lack ER- α (refs 12–14). All are close to seven feet tall and still growing, although it is not known whether their fertility is affected. Paradoxically, in some species, even male sexual/mating behaviour and ‘masculinization’ of the brain are oestrogen dependent¹⁵. Suddenly, the idea of ‘male’ and ‘female’ hormones begins to look thin. This is perhaps not surprising considering that most androgens can be converted into oestrogens (Fig. 1). The emerging concept is that androgens and oestrogens are the opposite sides of the same coin, and the balance in their actions may determine the biological response of target tissues.

One factor that has fuelled interest in oestrogen action in the male has been the furore over so-called ‘environmental oestrogens’. There is a possibility (still unproved) that exposure of the developing human male to such chemicals might have contributed to reported adverse changes (such as reduced sperm counts, genital anomalies and testicular cancer) in male reproductive health¹⁶. The main problem faced by those trying to

resolve this issue has been the lack of biochemical effects of oestrogens in the male. But this deficiency now looks set to be overwhelmed with the work of Hess *et al.*² and others. The huge, but exciting, task that we now face is to discover what oestrogens are doing at their many different sites of action in the male. □

Richard M. Sharpe is at the MRC Reproductive Biology Unit, 37 Chalmers Street, Edinburgh EH3 9EW, UK.

e-mail: r.sharpe@ed.rbu.mrc.ac.uk

1. Eddy, E. M. *et al.* *Endocrinology* **137**, 4796–4805 (1996).
2. Hess, R. A. *et al.* *Nature* **390**, 509–512 (1997).
3. Fisher, J. S. *et al.* *J. Endocrinol.* **153**, 485–495 (1997).
4. Hess, R. A. *et al.* *J. Androl.* (in the press).
5. Iljo, K. Y. & Hess, R. A. *Anat. Rec.* **234**, 190–200 (1992).
6. Chan, H. C. *et al.* *Biol. Reprod.* **52**, 192–198 (1995).
7. Brown, D., Verbavatz, J., Valenti, G., Lui, B. & Sabolic, I. *Eur. J. Cell Biol.* **61**, 264–273 (1993).
8. Janulis, L. *et al.* *J. Androl.* **17**, 111–116 (1996).
9. Hansen, L. A., Chulow, J. & Jones, R. C. *Int. J. Androl.* (in the press).
10. Saunders, P. T. K., Maguire, S. M., Gaughan, J. & Millar, M. R. *J. Endocrinol.* **154**, R13–R16 (1997).
11. Kuiper, G. G. J. M., Enmark, E., Peltö-Hukko, M., Nilsson, S. & Gustafsson, J.-A. *Proc. Natl Acad. Sci. USA* **93**, 5925–5930 (1996).
12. Carani, C. *et al.* *N. Engl. J. Med.* **337**, 91–95 (1997).
13. Smith, E. P. *et al.* *N. Engl. J. Med.* **331**, 1056–1061 (1994).
14. Morishima, A., Grumbach, M. M., Simpson, E. R., Fisher, C. & Qin, K. *J. Clin. Endocrinol. Metab.* **80**, 3689–3698 (1995).
15. Hutchison, J. B. *Curr. Opin. Neurobiol.* **1**, 562–570 (1991).
16. Toppari, J. *et al.* *Environ. Health Perspect.* **104**, S741–S803 (1996).

Magnetospheric physics

Big storms make little storms

George Siscoe

Magnetic storms bring auroras to sub-polar latitudes and perturb the magnetic field all over the Earth, disturbing navigational systems and sometimes even electrical power grids. They are caused by sudden changes in the currents flowing close to the Earth, particularly the ‘ring current’ that flows intermittently within the Earth’s magnetosphere (Fig. 1). What builds up the ring current, and thus causes storms? It was thought that the ring current was fed by smaller magnetic disturbances called ‘substorms’ — the huffs of a magnetic pump inflating the magnetosphere with hot plasma. But a consensus has emerged that substorms are instead earthquake-like releases of magnetic stresses, allowing the ring current to be built up by the larger-scale interplanetary magnetic field, which forces plasma from the magnetosphere’s tail into its interior.

Such is the proposal made early this October at a conference* called Brazil 5, by a small group of space physicists who meet biannually to discuss the causes and effects of magnetic storms. The group decided that new observational and computational findings are inconsistent with the older huff-and-puff picture.

The solar wind is a more remote cause of magnetic storms¹, because, in conjunction

with the Earth’s magnetic field, it forms the magnetosphere, and feeds the magnetotail with plasma. The question is, what transfers this plasma into the inner magnetosphere? Sydney Chapman gave substorms this role² in 1962. He thought substorms cause storms the way rain causes puddles: puddles happen when rain delivers water to shallows faster than it can evaporate; similarly, magnetic storms happen when substorms deliver hot plasma to the inner magnetosphere faster than it can dissipate.

This picture served space physics for 30 years. It was observed that substorms always accompany storms. It was found from linear prediction theory that the ‘AL’ index, which measures substorm activity, predicts 70% of the variance in the ‘Dst’ index, which measures storm intensity. (Both indices measure the amplitude of the magnetic disturbance at magnetic observatories around the globe. AL comes from observatories in the auroral regions and Dst comes from mid-latitude observatories.) Finally, it was found that substorms inject hot plasma from the tail into the magnetosphere.

So Chapman’s picture worked — until the beginning of this decade, when Yohsuke Kamide (Nagoya Univ.) asserted that “The main phase of magnetic storms develops because of sustained, southward interplanetary magnetic field (IMF), not because of

frequent intense substorms”³. A southward IMF, as it is carried along by the solar wind, grabs onto the outer layer of geomagnetic field within the magnetosphere and carries it along too. (By contrast, a northward IMF slips around the magnetosphere.) This sets up a general circulation within the magnetosphere, rather like the circulation within a raindrop set up by friction with the air through which it falls. This is the mechanism that Kamide credits with causing magnetic storms. The IMF usually points roughly in an east–west direction, and it takes a large solar disturbance, such as a coronal mass ejection, to upset the IMF’s normal orientation and swing it southward for long.

Kamide’s assertion sowed the seeds of a revolution that flowered about the time of a Chapman conference in 1996, when, ironically, Chapman’s picture of the substorm–storm connection came under heavy fire. Two studies^{4,5} showed that substorms slightly weaken storms. As these results loosened the traditional substorm–storm connection, the IMF–storm connection strengthened. In fact, the interplanetary magnetic field can be used to predict 76% of the variance of Dst and 61% of the variance in AL. This accounts for the predictive power of AL — both Dst and AL are governed by the IMF (ref. 5).

Meanwhile, a numerical simulation had been designed to test the Chapman picture,

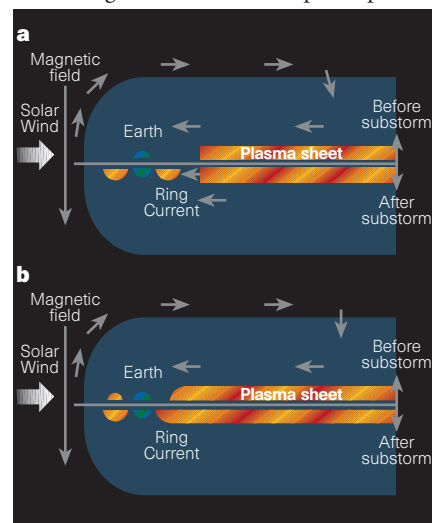


Figure 1 Magnetic weather — old and new pictures of the substorm–storm relationship. The Earth’s magnetic field forces the protons and electrons of the solar wind around the planet, carving out a comet-shaped volume called the magnetosphere. But some charged particles from the solar wind penetrate the magnetosphere, building up a region of plasma in the tail of the magnetosphere. In a magnetic storm, plasma moves inwards from the ‘geotail’, strengthening the ring current, a region of plasma around the Earth. In the old picture (a), substorms carried quantities of plasma from tail to ring; in the new picture (b), they release pressure, allowing the tail to spill over.

*Brazil 5: Workshop on Substorms–Storm Connection, Kreuth, Germany, 1–3 October 1997.

modelling the dynamics by which plasma moves from the tail into the magnetosphere. It showed that substorms have little effect on storm intensity, compared with IMF-driven plasma circulation (R. Wolf, Rice Univ.)⁶. This model had previously been used successfully to predict magnetic storm intensity, without the presence of substorms. So the Chapman conference produced a surprising convergence of evidence that the solar wind and the IMF cause storms. But what of substorms? Might they be mere by-products of the process, having no essential part to play?

A possible role for substorms, and a consolidation of the ground gained at the Chapman conference, were the main outcomes of Brazil 5. A plot of Dst averaged over many substorms showed no increase in storm intensity with a sudden increase in substorm activity (R. McPherron, UCLA). Nonetheless, substorms might play an essential supporting role. IMF-driven circulation produces high plasma pressures where plasma leaves the tail and enters the inner magnetosphere, which should choke the flow and stop the circulation⁷. Substorms may solve this pressure problem by releasing magnetic stresses, and so allow IMF-driven

circulation to complete its job (M. Hesse, Goddard Space Flight Center). This has implications for substorm prediction, because Dst is generally easier to predict than AL from solar wind and IMF measurements.

Brazil 5 replaced the traditional Chapman picture with the Kamide picture, but a version amended to give substorms an essential role. The term substorm, meaning a small version of a storm (as a raindrop is a subpuddle) is now something of a misnomer. A term in the Chapman mould (with a Latin or Greek prefix) more fitting to the Kamide picture would be 'synstorm', meaning 'together with'; but no one suggests changing names. □

George Siscoe is at the Center for Space Physics, Boston University, Boston, Massachusetts 02215, USA.

e-mail: siscoe@buasta.bu.edu

1. Gonzalez, W. D. *et al.* *J. Geophys. Res.* **99**, 5771–5792 (1994).
2. Chapman, S. *J. Phys. Soc. Jap.* **17**, 6–16 (1962).
3. Kamide, Y. *J. Geomag. Geoelectr.* **44**, 109–117 (1992).
4. Iyemori, T. & Rao, D. R. K. *Ann. Geophys.* **14**, 608–618 (1996).
5. McPherron, R. L. in *Geophys. Monogr.* **98** (eds Tsurutani, B. T. *et al.*) 131–147 (AGU, Washington DC, 1997).
6. Wolf, R. A. *et al.* in *Geophys. Monogr.* **98** (eds Tsurutani, B. T. *et al.*) 161–172 (AGU, Washington DC, 1997).
7. Erickson, G. M. & Wolf, R. A. *Geophys. Res. Lett.* **7**, 897–900 (1980).

Conservation biology

Poor prospects for oiled birds

Chris Mead

Two reports on the fate of seabirds cleaned and released after being caught in oil slicks make it clear that such rescue attempts largely fail. Both studies are of guillemots, *Uria aalge*, which are especially prone to oiling because they spend most of their time swimming on the sea surface. Wernham *et al.*¹ and Camphuysen *et al.*², who report on experience in Britain and the Netherlands, respectively found that only about 1% and 20% of the birds survived their first year after release. Healthy breeding guillemots have annual survival rates of over 90%¹.

Treatment of oiled birds varies from species to species, and with the type of oil³. Typically, in Britain oiled guillemots are thoroughly cleaned externally with a warm 2% solution of Fairy Liquid, a household washing-up liquid. This procedure takes two people 45 minutes or more. Waterproofing is restored by thorough removal of all oil residues using pressure rinsing with warm water. The birds are also given a rehydration mix including salt and glucose, with activated charcoal and fine kaolin, to purge ingested oil. They are then kept on a pool with surface drainage to remove even the thinnest sheen of oil. Provided their buoyancy and body weights are satisfactory, and the weather is good, birds may be released within a week of the initial incident.

Both studies^{1,2} involved ringing the birds on release after rehabilitation, and comparing the subsequent reports with those of healthy, wild birds ringed as part of a British national scheme. Wernham *et al.*¹ had 77 recoveries (subsequent reports) from 2,912 rehabilitated guillemots since 1985, which could be compared with 113 recoveries of full-grown wild birds out of 11,844. The median times survived were 7 days and 599 days, respectively. The best estimate was that 0.6% of the rehabilitated birds survived their first year — which is between 0.7% and 1.3% of what would be expected for healthy birds. There seems to be no significant difference between the survival of rehabilitated birds since 1985, when records started to be kept in more detail, and those ringed before then.

Camphuysen *et al.*² report 108 recoveries of 1,723 rehabilitated guillemots ringed in the Netherlands, mostly since 1988, and compared these figures with 2,176 birds ringed as adults at the colony on the Isle of May in the Firth of Forth, Scotland. The median recovery time for the former group was 12 days and for the Isle of May birds 671 days. The authors conclude that 22% of the rehabilitated birds survived their first year. But the only measure of ultimate success is for rehabilitated birds to rejoin the breeding population. Here at least there are signs of success, stemming from the work of Harris



100 YEARS AGO

As is well known, Chinese tea is mainly made from *Thea chinensis*, whilst that from India is the produce of *T. assamica*. Mr. Crole calls the former variety "a poor, scrubby-looking shrub" and "a wretched plant." From his remarks it would appear that the only service that Fortune rendered to the Indian tea industry by the introduction of the Chinese plant was the deterioration of the indigenous seed, giving rise to quantities of hybrids of various qualities, "from very rank stuff to fairly good." According to the author, the most that can be said in favour of the China plant is that it is distinctly more hardy than the Assam variety. We are disposed to believe that Mr. Crole's prejudices affect his judgement. There is no question whatever that some of the finest and most wholesome tea the world produces is to be met with in China. ... Mr Crole is of opinion that the value of a tea should be in direct proportion to the theine it contains. This is surely no more true than that the value of a wine depends upon its alcoholic strength. From *Nature* 2 December 1897.

50 YEARS AGO

Robert Liston (1794–1847), who died in his prime of an aortic aneurysm on December 7, 1847, is chiefly remembered as the first surgeon in Europe to operate under ether anaesthesia. He remarked at the time: "This Yankee dodge beats mesmerism hollow". He was not a good writer or speaker, and he contributed little to the science of surgery, but he was unsurpassed as a lightning and dexterous operator, whose methods of crushing stone and amputating thighs were the envy and despair of other surgeons.

'Cultivation of the cricket bat willow' – In the past, a number of private owners have tried to undertake the production of this high-quality willow timber, with the object of selling it as material for cricket bat production, and considerable failures have resulted. It is not that the cultivation of the willow itself is actually difficult, but a great deal of close attention is necessary, and more especially during its early life, and the right variety must be obtained at the start. It is for this reason that the Forestry Commission has had drawn up and issued this bulletin on the cultivation of this valuable product. From *Nature* 6 December 1947.

and Wanless⁴, who have thoroughly searched the ledges of part of the Isle of May colony for ringed birds. Over 1,000 of their own have been identified, together with 72 individuals from other colonies and four ringed outside Britain. One of these was not identifiable as an individual, but the other three were rehabilitated oiled birds, two from the Netherlands and one from Germany.

These results^{1,2} are the first from extensive and long-term data in Europe. In an analysis of records in the United States, however, Sharp⁵ came to the same conclusion for guillemots (median survival of 6 days for cleaned guillemots, as against 216 days for healthy birds), and for two other species: western grebe *Aechmophorus* sp. (11 and 624 days) and a sea-duck, the velvet scoter *Melanitta fusca* (7 and 466 days). Sharp was also able to assess the chances of oiled birds being cleaned and released for different incidents, and found that they varied widely (from 9% to 60%). Such data are not readily available from the European sources — given the chaotic circumstances when enthusiastic volunteers rally together in the face of an environmental disaster, it is impossible to find out how badly affected and what treatment birds that survived to release had undergone. However, Camphuysen (personal communication) considers that the Dutch birds had probably benefited from a longer period of recuperation than those from Britain, which may partly explain their better survival rates.

These results show that the best efforts to clean up oiled birds are not very successful. Nonetheless, bird welfare organizations would claim that it is their job to do what they can and that, even if only 1% survive, the work is worthwhile. Wernham *et al.*¹ also point out that there are happier cases. In Britain, mute swans (*Cygnus olor*) are often treated successfully, as are the penguins *Spheniscus demerus*, with up to 84% of rehabilitated birds later being seen in their South African colonies⁶.



Figure 1 What hope? A victim of the *Sea Empress* disaster.

Much may depend on the type of pollutant, for various sorts of oil have various effects on seabirds and waterfowl, and different treatments may be appropriate. The long-term toxic effect of oil ingested after preening from the plumage may be responsible for many deaths; other birds may not have regained fully effective waterproofing — the bird may perform well over a two- or three-hour test but not really be fit for continuous swimming. The oils involved in the incidents affecting *S. demerus* may account for the apparently better success rate with these birds — the oils are usually not as thick as those at higher latitudes, and they are quickly degraded as the warmer weather evaporates light fractions. Penguins also have particularly robust plumage.

What else can be done? The oil that causes the problem in the first place comes from illegal release or accidental spillage from ships. Better detection of covert tank cleaning operations, of the fuel tanks of ordinary vessels as well as tankers, should be an international priority (perhaps involving better use of satellite technology to identify and track the guilty ships). Recommendations for the improved management of accidents should come from the enquiry into the *Sea Empress* disaster, which occurred in South Wales in spring 1996; double-hulled tankers, the design of which contains oil after a collision, are also being phased in by international agreement.

The new results^{1,2} could be taken to mean that cleaning programmes are simply not worth it, and it is true that the conservation value is minimal for the guillemot population of more than a million birds in the North Sea. But while volunteers are prepared to put in the time and effort to try to save oiled birds, they will continue. Procedures could be improved. Accurate records of individual birds need to be kept from the moment they are found, so that the efficacy of different treatments can be properly assessed. Each bird thought to be suitable for release should then be identified with long-lasting metal rings, and annual ringing of wild, healthy birds should also continue — in part for comparative purposes, in part to identify the colonies involved in the oiling incidents that will sadly continue to occur. □

Chris Mead is at The Nunnery, Hilborough, Thetford, Norfolk IP26 5BW, UK.
e-mail: chris.mead@zetnet.co.uk

1. Wernham, C. V., Peach, W. J. & Browne, S. J. *British Trust for Ornithology Res. Rep.* 186 (Thetford, Norfolk, 1997).
2. Camphuysen, C. J., Duiven, P., Harris, M. P. & Leopold, M. F. *Sula* 11, 157–174 (1997).
3. Robinson, I. in *BSAVA Manual of Raptors, Pigeons and Waterfowl* (eds Beynon, P. H., Forbes, N. A. & Harcourt-Brown, N. H.) Ch. 35 (Br. Small Anim. Vet. Assoc., Cheltenham, 1996).
4. Harris, M. P. & Wanless, S. *Sula* 11, 183–185 (1997).
5. Sharp, B. E. *Ibis* 138, 222–228 (1996).
6. Underhill, L. G., Whittington, P. A., Crawford, R. J. M. & Williams, A. J. *Sula* 11, 187–192 (1997).

Daedalus

The ultimate Sun-block

A solar eclipse, the passage of the Moon across the face of the Sun, is a rare and exciting event. The Moon's shadow brings a sort of momentary night; the blue of the sky vanishes as sunlight is no longer scattered by the atmosphere; the stars come out. Daedalus now suggests a way of making an artificial solar eclipse.

His idea is to obstruct the Sun's disk with a high-flying circular shutter. To block most of the blue atmospheric scattering above a given point, it needs to be about 20 km across. It would subtend the Sun's diameter at some 2,000 km altitude — its upper limit of height. Clearly, it has to be in orbit. Daedalus recalls the scheme for a 'solar sail' spacecraft deploying a vast sail of aluminized polymer film. The pressure of solar radiation on its sail would propel it. His 'Eclipsat' will draw on much the same technology.

Eclipsat will resemble the inner tube for a bicycle tyre 20 km across, folded up small and immersed in a special viscous monomer fluid. Released in orbit, the toroidal tube will inflate under its internal pressure, and will slowly unfold into true circular form. The viscous monomer will be spread out like a soap-film as the tube unfolds, ultimately forming a perfect disk within it, 20 km across and a few micrometres thick. Solar radiation will soon polymerize it to a solid film. Sadly, no dye in the polymer could make such a thin film opaque. It will have to be coated with metal, possibly from a nearby pyrotechnic evaporator released from the same rocket.

The ideal orbital height for Eclipsat is perhaps 1,000 km. It will then produce a solar eclipse every 105 minutes along a track about 5 km across and maybe 8,000 km long. At any point on the track, each eclipse will last only 2 seconds; but their steady repetition will soon yield detailed information on solar prominences and their evolution with time, starlight shifts, and so on. Observers will see a giant shadow racing towards them at about 7 km per s, engulfing them in a brief night.

At other points of its orbit, the metallized disk of Eclipsat will act as a mirror. It will reflect the Sun down onto a narrow track on the dark side of the Earth, giving the novel phenomenon of an 'anti-eclipse'. Observers will see a band of brightness racing through the night towards them; for 2 seconds it will be bright, dazzling, sunlit day; then darkness will descend again. It will be an awe-inspiring spectacle.

David Jones

Albers's abstracts

Josef Albers aimed to neutralize all elements bar one in his paintings, leaving colour as the only variable. The concept has strong parallels with scientific experiments to test theories.

Martin Kemp

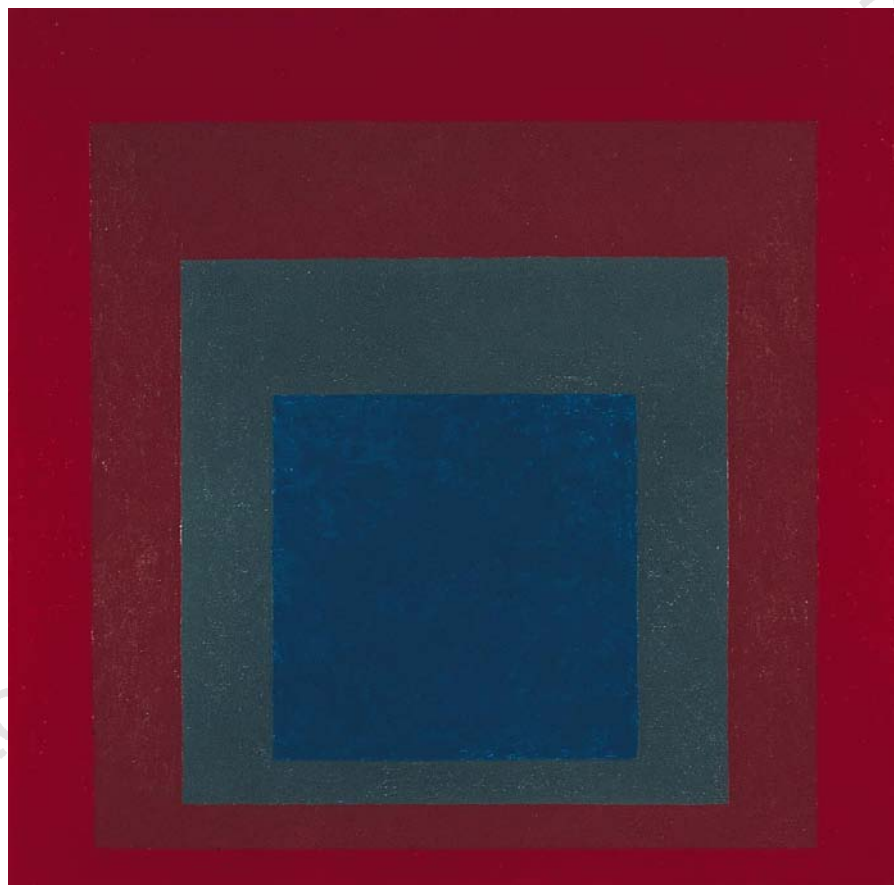
The effects of colour in painting would, on the face of it, seem to be one of the most obvious areas in which to forge a 'science of art'. Yet the colourist's science remains, in the words of a sixteenth-century commentator, the "true alchemy of painting". The colourist-magician compounds base materials into radiant wholes that invariably seem to transcend systematic explanation and predictability.

The old Aristotelian system of colour, which ranged hues along a scale from lightness to darkness, actually worked quite well with painters' practice. A yellow pigment, for example, inherently occupied a position at the light end of the tonal scale, while ultramarine blue stood closer to black. Far from clarifying painters' practice, Newton's spectral division of white light presented sustained problems, not least before Hermann von Helmholtz's elucidation of the essential difference between subtractive colour mixing for pigments and additive mixing for lights.

The key problem for the colourist arises from the high levels of fluidity in our perception and subjective response to the same colours in different combinations and configurations, on various scales and in variable conditions of viewing.

This fluidity continues to provide complex challenges for experimental psychology. A logical response is for the painter to adopt an empirical stance, giving primacy to practical testing rather than theoretical prediction. This is exactly the course followed by Josef Albers in his series *Homage to the Square* and his book *Interaction of Colour* in 1963.

As a student and teacher at the Bauhaus in



Josef Albers's *Homage to the Square* (Haags Gemeentemuseum, The Netherlands).

Germany from 1920 to 1933, where he worked directly with Paul Klee, Albers reached maturity in the environment of experimental questing that characterized this extraordinary modernist laboratory for art and architecture.

It was in 1950, while working at Yale University, that he arrived at his definitive fusion

of practice and theory. The square was intended as a neutral container — as "the dish I serve my craziness about colour in". The matt areas of colour similarly aspire to establish a neutrality devoid of the personal handwriting and individual style of traditional painting. The resulting vehicle is akin to a modern scientific experiment in which just one variable — in this case colour — is left unfixed. The 'experiments' and results are then set beside the tenets of colour theories, such as those of Goethe and Wilhelm Ostwald, the Nobel prize-winning physical chemist who published *The Colour Primer* in 1916. The science of colour, for Albers, can inform practice but never provide absolute prescriptions for the making of works of art.

Each work may be regarded as an experiment that is simultaneously dependent on the laws of colour and individually subversive of their predictive power. The elusive complexity of the particular arises from visual means of the utmost simplicity. □

Martin Kemp is in the Department of the History of Art, University of Oxford, 35 Beaumont Street, Oxford OX1 2PG, UK.



Albers's *Section Through the Colour Solid of Wilhelm Ostwald* from *Interaction of Colour* (1963).

Caterpillar kinematics

The design of a caterpillar's body forces it to walk slowly — but even though it cannot run away from danger, it can roll up quickly into a protective coil. Here, I show that the mother-of-pearl moth, *Pleurotya ruralis*, uses this reflex coiling as the basis for a method of high-speed escape. By anchoring the end of its body to the ground and recoiling against it, the larva converts itself into a backwardly rolling wheel. In doing this, it shows that the limitations of a soft, segmented body can be overcome, using the basic assemblage of segmental muscles, by temporarily sacrificing the need for stability.

During forwards locomotion a wave of contraction followed by relaxation runs along the body from tail to head, producing the characteristic travelling 'hump'¹⁻³ on a caterpillar's back. As the wave passes, each segment is raised from the ground, is telescoped forwards into its neighbour in front, then is lowered back to the ground (Fig. 1a). The legs attached to terminal segment 13 (claspers) and segments 6-9 are retracted, borne forwards with their segment, then put down again one 'step' ahead. Hence for the whole body to progress forwards one step, each segment needs to take its own step forwards. Each foot is airborne for only 35% of the cycle: although this maximizes stability it prevents the body from building up significant momentum. Caterpillar crawling is thus slow and uneconomical compared with terrestrial locomotion in animals with muscles harnessed to a rigid

skeleton. Caterpillars walk at only 10% of the maximum speeds seen in adult running insects of the same weight⁴⁻⁷.

Like other moth larvae that seek protection inside rolled-up nettle leaves, the larva of *Pleurotya ruralis* lacks most of the normal caterpillar defences against birds and ichneumon wasps, such as cryptic coloration, irritant hairs or unpalatability⁸. But if an individual is isolated on a flat surface and provoked by mechanical shock, it displays a range of withdrawal responses of increasing speed and intensity, all based on locomotion in reverse.

Mild shock to the head or thorax elicits backwards walking, the exact kinematic inverse of walking forwards in that the peristaltic wave now begins at the head and moves tailwards. Stronger stimulation provokes a much faster wave, which arches up the whole body, wrenching all the legs free of the ground except the terminal claspers (Fig. 1b). The latter now serve as an anchor to which the body can be jerked during its 'airborne' phase, until the intervention of the relaxation phase of the wave lowers the legs to the ground. As the relaxation wave reaches the claspers they detach, then reattach further back, thereby completing the step. Technically, this fast running retreat can be considered a 'reverse gallop' because the entire body (except the claspers) is airborne for a significant part of the locomotory cycle.

The third and most dramatic escape response begins like a reverse gallop but this

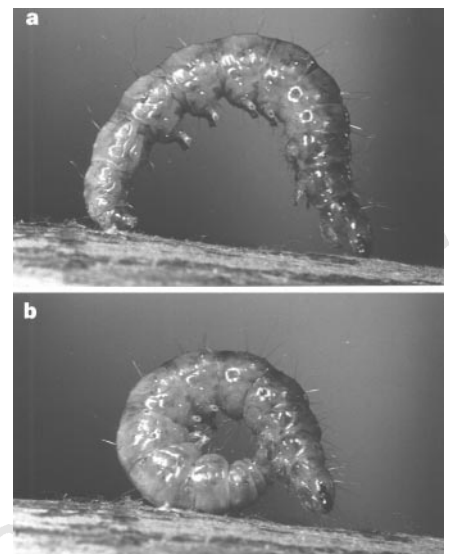


Figure 2 Still photographs of two stages in the recoil-and-roll manoeuvre of *Pleurotya* caterpillar. **a**, Recoil; **b**, beginning of roll, with claspers detached from ground. Flash exposures at 1/8000 s.

time the relaxation wave fails to appear. The body continues to move under its own momentum, coils up into a wheel, and begins to roll backwards releasing the claspers as it goes (Figs 1c and 2). Depending on initial conditions, such as the flatness and texture of the surface, and any slight imbalances in the body during arch formation, the momentum is sufficient to produce up to five complete revolutions, and a withdrawal speed of $39 \pm 3.6 \text{ cm s}^{-1}$ ($n = 31$), nearly 40 times normal walking speed.

Recoil-and-roll uses the same set of segmental muscles as ordinary locomotion, but more quickly, dispensing with the need for serial contraction. The manoeuvre is derived from the basic coiling response of caterpillars under attack from their natural enemies. By harnessing the output of the flexural muscles involved in spiralling to a fixed point provided by the claspers, the caterpillar converts the coil into a rolling wheel. I have found no evidence for the involvement of an elastic spring in this mechanism, such as the pre-tensioning of the skin that allows maggots to leap up to 8 cm into the air⁹; on the contrary, recoil-and-roll makes maximum use of the power available from real-time muscle contraction.

John Brackenbury

Department of Anatomy, Downing Street,
Cambridge CB2 3DY, UK

e-mail: jhb@mole.bio.cam.ac.uk

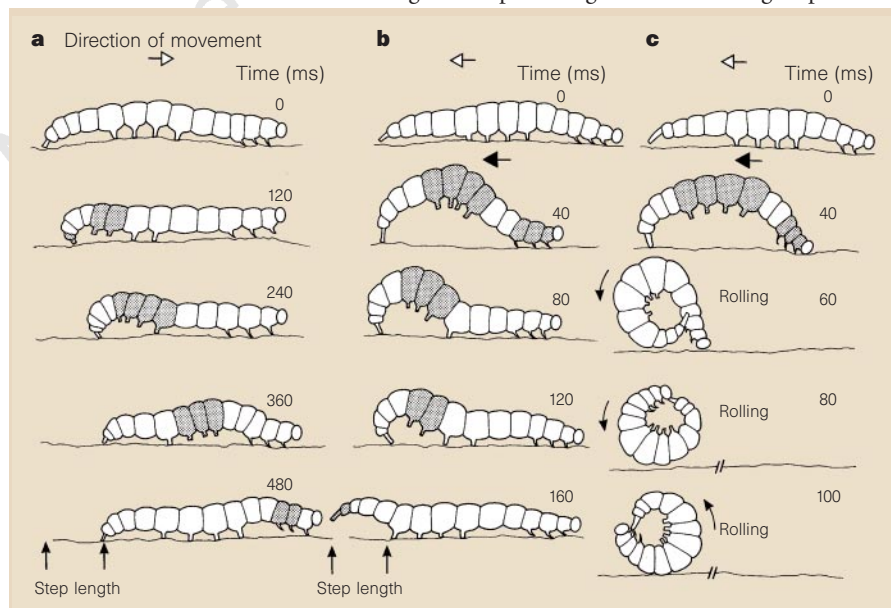


Figure 1 Main locomotory gaits in *Pleurotya* caterpillar. In **a** the insect is walking from left to right; in **b** and **c** it is retreating from right to left. Stippling indicates when leg-bearing segments (6-9, 13) are off the ground. The black arrow in **b** and **c** signifies the rapid recoil stage of locomotion. Speed, stride frequency and stride length during normal forward walking were $1.0 \pm 0.2 \text{ cm s}^{-1}$, $1.7 \pm 0.2 \text{ Hz}$ and 0.6 cm , respectively. Measurements were made from video images (Panasonic, 50 frames per second, shutter speed 0.001 s, and NAC200, 200 frames per second with strobe synchronization).

1. Barth, R. *Zool. Jb.* **62**, 507-566 (1937).
2. Hughes, G. M. & Mill, P. J. in *The Physiology of Insects* Vol. 3 (ed. Rockstein, M.) 335-379 (Academic, New York, 1974).
3. Brackenbury, J. H. *Physiol. Entomol.* **21**, 7-14 (1996).
4. Casey, J. M. *Science* **252**, 112-114 (1991).
5. Joos, B. *Physiol. Zool.* **65**, 1148-1161 (1992).
6. Berrigan, D. & Lighton, J. R. B. *J. Exp. Biol.* **179**, 245-259 (1993).
7. Berrigan, D. & Pepin, D. J. *J. Insect Physiol.* **41**, 329-337 (1995).
8. Edmunds, M. *Defence in Animals* (Longman, London, 1974).
9. Maitland, D. P. *Nature* **355**, 159-161 (1992).

No 'nanofossils' in martian meteorite

Elongated, segmented forms found on fracture surfaces within the martian meteorite ALH84001 have been proposed to be martian 'nanofossils'¹, even though they appear too small to be fossilized bacteria²⁻⁴. We have examined similar forms and find that the majority are (non-biological) lamellar growth steps on pyroxene and carbonate crystals. Their segmented surface microstructures are laboratory artefacts resulting from the deposition of conductive heavy-metal coatings.

Using transmission electron microscope (TEM) imaging, the definitive method for identifying nanofossils⁵, we have previously observed magnetite whiskers but no nanofossils on or near the surfaces of the carbonates⁶. We suggested that some of the purported nanofossils¹ might be similar high-temperature, vapour-deposited magnetite whiskers, but the dissimilarity between specimen preparation and electron-beam analysis techniques used in the two studies makes direct comparisons difficult.

We have now examined fracture surfaces from carbonate-rich regions using the techniques described by McKay *et al.*¹. Chips of ALH84001 with exposed carbonate and pyroxene crystals were mounted, coated

with 2–20 nm of gold or gold/palladium, and examined in a field emission scanning electron microscope (FE-SEM) using secondary electron imaging and incident electron beam energies of 2–20 keV. We measured mineral compositions using energy dispersive X-ray spectroscopy (EDS).

The surface topography is highly irregular on a nanometre scale, with emergent lamellae following the major cleavage direction of the substrate (Fig. 1a). Some regions exhibit typical pyroxene cleavage, whereas others display lamellar features (Fig. 1b). Close examination of these features over a range of tilt angles reveals that they are angled directly into the pyroxene substrate and are the emergent edges of substrate lamellae. Many of these lamellae have segmented, curvilinear surface microstructures resembling the elongated forms proposed to be nanofossils by McKay *et al.*¹.

The surfaces of carbonates within the fracture zones of ALH84001 are also decorated with emergent lamellae (Fig. 1c, d). Although we suspect that some are magnetite whiskers (composition and crystal structures of magnetite cannot be established by SEM-EDS), their abundance far exceeds that of the whiskers that we described previously⁶. As with the pyroxenes, the lamellae are angled directly into the carbonate substrate and their orientations are controlled by local substrate cleavage directions.

The segmented surface microstructures

of the elongated forms in ALH84001 (Fig. 1a, b) are largely an artefact of the conductive metal coating. Similar coating artefacts have previously been misidentified as (terrestrial) nanofossils⁷. Au and Au/Pd coatings produce segmentation on elongated forms, although Au/Pd imparts a finer-scale segmentation (as seen in the images in refs 1 and 8). In addition, the segmentation becomes more pronounced as the coating thickness is increased⁷ (compare Fig. 1a and b). As noted by McKay *et al.*¹, specimen charging is a problem with some chips of ALH84001, because the minerals are insulators and because some chips are shattered (shock-fractured) on a submicrometre scale. A conductive coating is desirable for secondary electron imaging, and the optimum coating (composition and thickness) is a trade-off between charging effects, surface artefacts (enhanced segmentation) and required image resolution.

The complexity of the pyroxene and carbonate fracture surfaces in ALH84001 probably reflects the superposition of severe shock and/or thermal events on indigenous surface microstructures of the minerals^{9,10}. Many of the lamellae relate to internal cleavage planes within the pyroxene and carbonate substrates, and their orientations may reflect localized shock deformation, partial melting, or even the onset of secondary alteration (for example, clay formation¹¹). Where lamellae are exposed on fracture surfaces, their enhanced secondary electron yield (a characteristic of lamellar surfaces^{12,13}) and segmented surface structures (an artefact of the metal coating) combine to produce the nanofossil-like appearance.

Although some of the elongated forms of ALH84001 could conceivably be martian nanofossils, the majority appear to be either emergent substrate lamellae or magnetite whiskers⁶. Some lamellae may exhibit indigenous segmentation¹⁴, but conductive metal coatings both produce and accentuate segmentation which is particularly prominent in high-resolution FE-SEM images. TEM imaging is therefore more appropriate for distinguishing coated lamellae from 'nanofossils', as it is possible to observe internal microstructures characteristic of non-biogenic⁶ and biogenic⁵ 'elongated forms'.

J. P. Bradley

MVA Inc., 5500/200 Oakbrook Parkway,
Norcross, Georgia 30093, USA
and School of Materials Science and Engineering,
Georgia Institute of Technology,
Atlanta, Georgia 30332-0245, USA

R. P. Harvey

Department of Geological Sciences,
Case Western Reserve University, Cleveland,
Ohio 44106-7216, USA

H. Y. McSweeney Jr

Department of Geological Sciences,
University of Tennessee, Knoxville,
Tennessee 37996-1410, USA

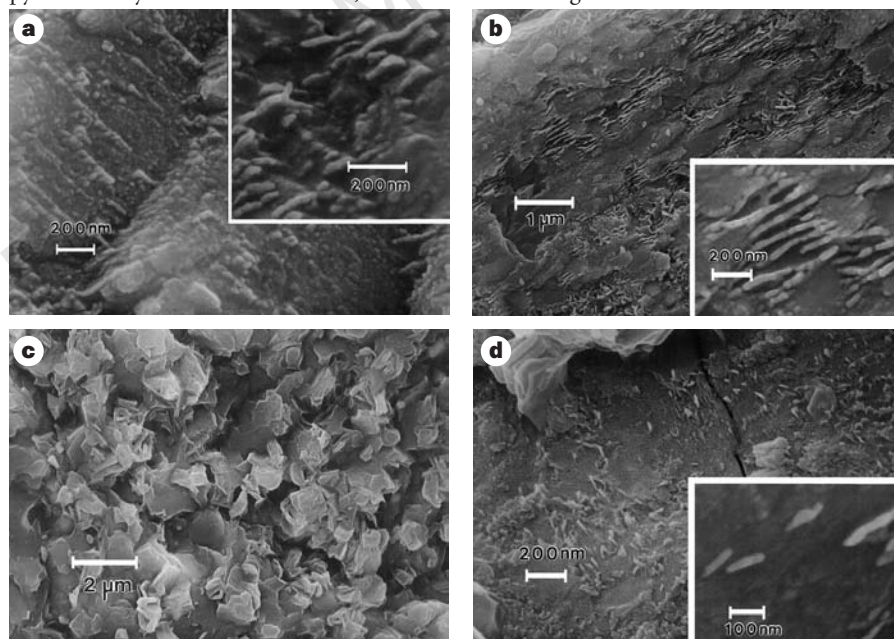


Figure 1 'Elongated forms' in the meteorite ALH84001. **a**, Secondary electron image of pyroxene surface. Parallel elongated forms (from upper left to lower right) are emergent lamellae related to substrate cleavage direction. Inset, pyroxene surface oriented such that elongated forms resemble nanofossils. Both surfaces have a 10-nm-thick coating of gold, measured by TEM imaging of ultramicrotomed cross-sections. **b**, Secondary electron images of highly fractured pyroxene surface. Inset, high-magnification image of lamellae. Coating: 20 nm Au, resulting in more pronounced segmentation. **c**, Secondary electron image of carbonate crystals (coated with 9 nm of Au/Pd). The average grain size is <1 μm as suggested by previous work⁹. **d**, Secondary electron image of carbonate surface within fracture zone. Inset, high-magnification image of elongated forms on carbonate, some of which may be magnetite whiskers⁶. Coating: 10 nm Au.

McKay *et al.* reply—We are aware that some mineral surfaces of ALH84001 display very narrow elongated forms, often subparallel, in the nanometre size range (Fig. 2a, b here and Fig. 1a, b and d of Bradley *et al.*). They are very common on pyroxene surfaces, but also occur on carbonate surfaces. We have also imaged some of these features, which Bradley *et al.* propose are simply emergent edges of pyroxene lamellae, with a field emission scanning electron microscope (FE-SEM). We have previously reported the presence of small amounts of clay minerals in ALH84001 from TEM images and electron diffraction data¹¹. We propose that the features in Fig. 2a, b here and Fig. 1b of Bradley *et al.* may be weathering features caused by incipient formation of clay minerals. In places, these figures have a 'platy' appearance characteristic of some clay minerals especially in the uncoated ALH84001 sample (Fig. 2a). Regardless of their exact origin, it is unlikely that these features are biogenic.

By contrast, the features that we imaged previously (Fig 6b in ref. 1) are not subparallel. They display intersecting alignment, have significant curvature and are more isolated. In addition, features tentatively interpreted as martian nanofossils are larger than the lamellae or elongated magnetite reported by Bradley *et al.*⁶. Curved and S-shaped features from the surfaces of carbonate globules are up to 0.75 μm in length (Fig. 2c), an order of magnitude larger than the 'lamellae' or the elongated magnetite observed by Bradley *et al.*⁶. Additionally, some of the features we proposed as possible martian microfossils are ovoid rather than elongated (Fig. 6a in ref. 1).

It is well known that conductive coatings can alter appearance, cover fine details and produce artefacts. We ran a series of controls for coating artefacts (ref. 50 in ref. 1). Carbon coating of ALH84001 samples obscures fine honeycomb-like surface texture and many of the larger features become rounded or softened (Fig. 2d, e). However, the dimensions of the larger features are not appreciably affected (Fig. 2d, e).

Previous studies on latex spheres¹⁵ show that coatings of Au produce larger artefact grains than Au/Pd; grain size of Au is 2–3 times larger than Au/Pd¹⁶. Bradley *et al.* used Au in producing their high-magnification images, so the purported artefact segmentation would naturally be more prominent than if Au/Pd were used. We used Au/Pd, carbon, or completely uncoated samples in our studies, never gold alone, and have also made control coatings on lunar glass surfaces using 30 and 60 seconds (~10 and 20 nm thick, respectively) of Au/Pd from our coater. The glass is not exactly analogous to the minerals in ALH84001 (except perhaps for the feldspathic glass), and artefacts are substrate-

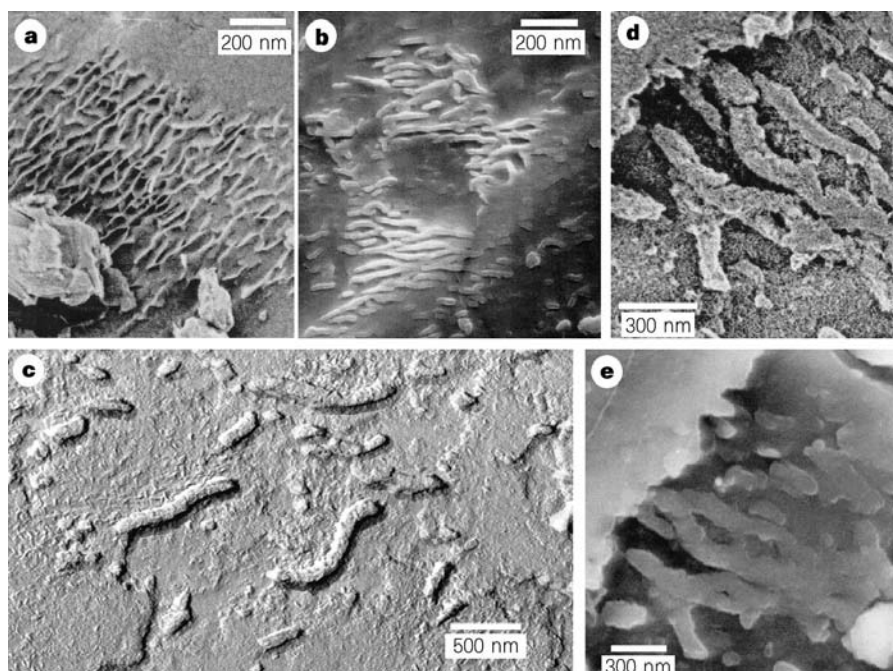


Figure 2 FE-SEM images of ALH84001 samples. **a**, Fine-scale platy texture on a carbonate surface in ALH84001. The sample was imaged at 1 kV without any conductive coating. **b**, The same surface after coating with a thin (~10nm) carbon layer. **c**, TEM image of a shadowed replica of part of a carbonate surface from ALH84001. The replica was made from an uncoated surface so textures in the image represent true textures of the surface. These elongate features are much larger than either the platy structures in (**a**) or the 'lamellae' and whisker magnetite described by Bradley *et al.*⁶. **d**, Texture on the surface of an uncoated carbonate in ALH84001 imaged at 1 kV. Note the fine-grained wispy texture which we interpret as an extremely thin clay mineral alteration of the carbonate. **e**, The same sample now coated with ~10 nm of carbon and imaged at 10 kV. Note that all of the large strand features are visible and relatively unchanged although space between them has been partly filled. The fine wispy surface now looks smooth.

dependent, but its smooth surface is ideal for detecting coating artefacts. At 30 s there are no artefacts and at 60 s only a fine uniform cracking or crazing texture is apparent. For our original ALH84001 studies, we generally used 30 s or less (and never more than 60 s) for sample coating. Although some of our images may have a fine crazing resulting from the coating, we have not observed major decoration in any of our control samples. ALH84001 substrates may accentuate and amplify coating artefacts more than our controls, but we have no evidence for such anomalous behaviour.

Bradley *et al.* suggest that the purported martian nanofossils are too small for fossilized bacteria. In fact, the lower size limit of fossilized bacteria has not been determined. In mammalian blood, living bacteria as small as 70 nm in diameter have been found which contain identifiable DNA and can reproduce^{17,18}. Also, soil bacteria as small as 80 nm have been found¹⁹.

Bradley *et al.* suggested that the purported elongated microfossils may be high-temperature, vapour-formed magnetite whiskers, but other studies have shown that the carbonate globules and their included magnetites could not have formed at temperatures above 100–300 °C^{20–22} whereas elongate magnetite grains can be produced by bacteria at low temperatures^{23,24}. They

also assert that TEM imaging is more appropriate for identifying nanofossils, as it is possible to distinguish internal microstructures. SEM and TEM imaging are complementary and the use of both methods on the same samples would provide the best evidence for biogenic activity, particularly when internal structure is destroyed or replaced by mineralization^{25,26}.

David S. McKay, Everett Gibson Jr

NASA Johnson Space Center,

Houston, Texas 77058, USA

e-mail: david.s.mckay1@jsc.nasa.gov

Kathie Thomas-Keprta

Lockheed Martin, 2400 NASA Road 1, C-23,

Houston, Texas 77058, USA

H. Vali

McGill University, 3450 University Street,

Montreal, Quebec H3A 2A7, Canada

1. McKay, D. S. *et al.* *Science* **273**, 924–930 (1996).

2. Maniloff, J. *Science* **276**, 1776 (1997).

3. Nealson, K. H. *Science* **276**, 1776 (1997).

4. Psenner, R. & Loferer, M. *Science* **276**, 1776–1777 (1997).

5. Bazylinski, D. A. *ASM News* **61**, 337–340 (1975).

6. Bradley, J. P., Harvey, R. & McSweeney, H. Y. Jr *Geochim. Cosmochim. Acta* **60**, 475–481 (1996).

7. Folk, R. L. *et al.* *J. Sedim. Res.* **67**, 583–587 (1997).

8. Kerr, R. A. *Science* **276**, 30–31 (1997).

9. Harvey, R. P. & McSweeney, H. Y. Jr *Nature* **382**, 49–51 (1996).

10. Scott, E. R. D., Yamaguchi, A. & Krot, A. N. *Nature* **387**, 377–379 (1997).

11. Thomas-Keprta, K. L. *et al.* *Lunar Planet. Sci.* **28**, 1433–1434 (1997).

12. Goldstein, J. I. & Yakowitz, H. (eds) *Practical Scanning Electron Microscopy* (Plenum, New York, 1975).

13. Brooks, C. R. & Bogni, F. *Mater. Character.* **38**, 103–117 (1997).
14. Steele, A. *et al.* *Lunar Planet. Sci.* **28**, 1369–1370 (1997).
15. Morrison, D. A. & Clanton, U. S. *Proc. Lunar Planet. Sci. Conf.* **10**, 1649–1663 (1979).
16. Newbury, D. E., Joy, D. C., Echlin, P., Fiori, C. E. & Goldstein, J. I. in *Advanced Scanning Electron Microscopy and X-Ray Microanalysis* (Plenum, New York, 1986).
17. Kajander, E. O., Kuronen, I., Åkerman, K., Pelttari, A. & Çiftçioglu, N. in *Proc. Soc. Photo-Opt. Instrum. Eng.* Vol. 3111 (ed. Hoover, R. B.) 420–428 (Int. Soc. Opt. Eng., Bellingham, WA 1997).
18. Çiftçioglu, N., Pelttari, A. & Kajander, E. O. in *Proc. Soc. Photo-Opt. Instrum. Eng.* Vol. 3111 (ed. Hoover, R. B.) 429–435 (Int. Soc. Opt. Eng., Bellingham, WA 1997).
19. Kieft T. L. in *Nonculturable Microorganisms in the Environment* (eds Colwell, R. R. & Grimes, D. J.) (Chapman and Hall, New York, in the press).
20. Romanek, C. S. *et al.* *Nature* **372**, 655–657 (1994).
21. Valley, J. W. *et al.* *Science* **275**, 1633–1638 (1997).
22. Gleason, J. D., Kring, D. A., Hill, D. H. & Boynton, W. V. *Geochim. Cosmochim. Acta* **61**, 3503–3512 (1997).
23. Hanzlik M., Winkhofer M. & Petersen N. *Earth Planet. Sci. Lett.* **145**, 125–134 (1996).
24. von Döbeneck T., Petersen N. & Vali H. *Geowiss. Uns. Zeit* **1**, 27–35 (1987).
25. Westall, F. in *Astronomical and Biochemical Origins and the Search for Life in the Universe* (eds Cosmovici, C.B., Bowyer, S. & Werthimer, D.) 491–504 (Compositori Editrice, Bologna, 1997).
26. Westall F. *Darmstädter Beiträge zur Naturges.* **4**, 29–43 (1994).

Travelling waves in vole population dynamics

Spatial self-organization patterns in population dynamics have been anticipated^{1–3}, but demonstrating their existence requires sampling over long periods of time at a range of sites. Voles cause severe economic damage and are therefore extensively

monitored, providing a source of the required data. Using two long-term data sets^{4–6} we now report the existence of travelling waves in vole population numbers.

Information about the intensity of annual vole damage to young forests was available for the years 1972–83 for 19 forestry board districts of Finland^{5,6}. The geographical distribution of the most damaged areas fluctuates (Fig. 1a, b) and the timing of damage is asynchronous among the different areas. Locally, the damage peaks at intervals of three to four years, matching the cyclic dynamics of Scandinavian voles⁷. This is in agreement with simulation studies² showing that spatial order is associated with the presence of spatial waves showing a certain degree of periodicity.

Correlation between the latitude of the 'centre of mass' and the annual severity of the damage (Fig. 1a, b) is $r=0.84$ ($P<0.001$; for the longitude $r=0.13$). Thus, with increasing volume of vole damage the damage epicentre moves north. There was also a negative correlation between intensity of damage and distance between provinces ($r=-0.47$; $P<0.001$) suggesting that nearby localities have similar damage levels, and providing another indication of spatial structure⁸ in vole damage.

We also analysed records of vole plagues in 53 departments of France for the years 1900–35¹ (Fig. 1c, d). There were rapid changes in the extent of the plagued area, but there were no clear periodicities

like those seen in the Scandinavian voles⁷. Again there was a correlation between the latitude of the annual location of the centre of mass and the geographical extent of the damage ($r=-0.67$, $P<0.001$; for the longitude $r=0.06$). Accordingly, the French vole plagues spread from north to south, whereas in Finland they spread from south to north.

A variable is said to be spatially autocorrelated when it is possible to predict the values of this variable in one site from the known values at nearby sampling sites⁹. We performed autocorrelation analyses for irregularly spaced interval (Finland) or nominal scales (France)¹⁰.

For the Finnish data, we computed annual autocorrelation coefficients^{9–11} for the distance class of 200 km assuming binary links. We detected significant positive spatial autocorrelation in six of the twelve years. For the French data we computed annual spatial autocorrelations using binary weights and assuming a link between sites that were less than 120 km apart. We found a significant spatial autocorrelation present in the data in 14 of the 33 years. Both results indicate that spatial structure exists in these populations.

By identifying spatial autocorrelation in vole populations and an annually moving epicentre of vole plagues as a function of damage extent, we conclude that travelling waves, or pulses, in the dynamics of vole populations exist, a phenomenon predicted^{1–3} by theoretical population ecology.

Esa Ranta

Department of Ecology and Systematics,
University of Helsinki,
PO Box 17,
FIN-00014 Helsinki,
Finland
e-mail: esa.ranta@helsinki.fi

Veijo Kaitala

Department of Biological and Environmental
Science,
University of Jyväskylä,
Box 35,
FIN-40351 Jyväskylä,
Finland

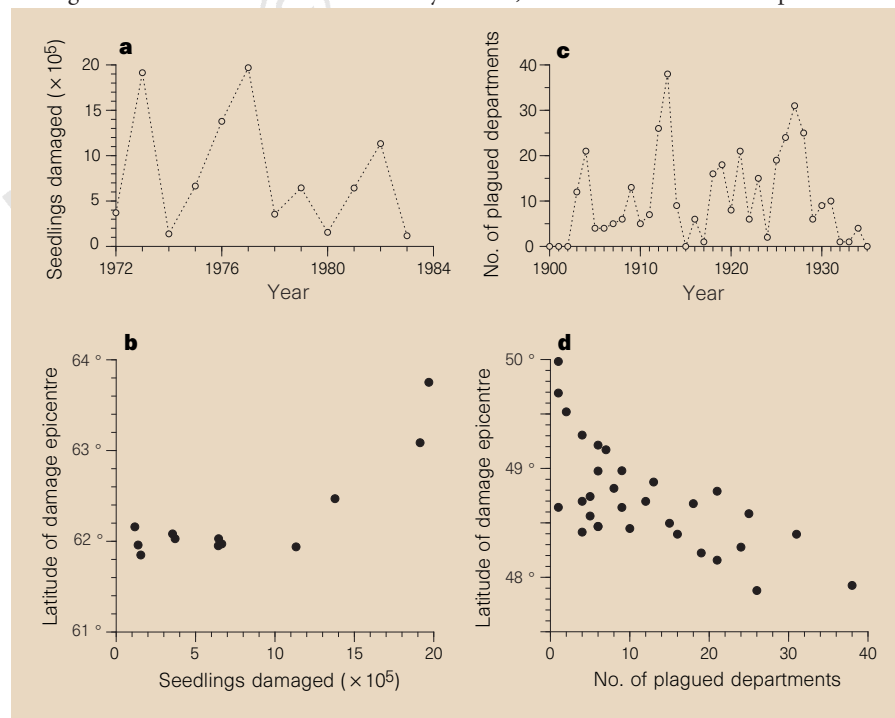


Figure 1 Extent and severity of vole damage. **a**, The number of seedlings destroyed by voles on study plots in young forest plantations in Finland^{5,6}. **b**, The dynamics are characterized by movement of the annual damage epicentre from south to north as total damage increases. **c**, Geographical extent of French⁴ vole plagues; and **d**, annual epicentre of plagues plotted against the number of plagued departments.

1. Hassell, M. P., Comins, H. N. & May, R. M. *Nature* **353**, 255–258 (1991).
2. Rohani, P. & Miramontes, O. *Proc. R. Soc. Lond. B* **260**, 335–342 (1995).
3. Bascompte, J. & Sole, R. *Trends Ecol. Evol.* **10**, 361–366 (1995).
4. Elton, C. *Voles, Mice and Lemmings. Problems in Population Dynamics* (Clarendon, Oxford, 1942).
5. Teivainen, T. *Folia Forestalia* **387**, 1–23 (1979).
6. Teivainen, T. *Metsäntutkimuslaitoksen Tiedonantoja* **145**, 1–12 (1984).
7. Hansson, L. & Henttonen, H. *Oecologia* **67**, 394–402 (1985).
8. Ranta, E., Kaitala, V., Lindström, J. & Lindén, H. *Proc. R. Soc. Lond. B* **262**, 113–118 (1995).
9. Sokal, R. R. & Oden, F. M. L. S. *Biol. J. Linn. Soc.* **10**, 199–228 (1978).
10. Legendre, P. & Fortin, M.-J. *Vegetatio* **80**, 107–138 (1989).
11. Moran, P. A. P. *Biometrika* **37**, 17–23 (1950).

Second careers of the Nazis' doctors

Kontinuität und Neuanfang in der Hochschulmedizin nach 1945*

edited by Gerhard Aumüller, Hans Lauer and Helmut Remschmidt
Schüren: 1997. Pp. 128. DM18 (pbk)

Marco Finetti

During the Third Reich (1933–45), Otmar von Verschuer was one of the principal representatives of the so-called *Rassenhygiene*, the Nazis' genetics and eugenics programme. As a professor at Frankfurt University and later director of the renowned Kaiser-Wilhelm Institute of Anthropology, Verschuer and his colleagues provided the 'scientific' basis for Nazi race laws, mass sterilizations and acts of euthanasia to which thousands fell victim. Verschuer taught SS doctors and also maintained close contacts with Josef Mengele, whom he knew as a doctoral student at Frankfurt University. This association with the Auschwitz doctor was the main reason for Verschuer's dismissal and removal from office after the collapse of the Nazi regime in 1945.

But his career was by no means over. In 1951, Verschuer was appointed professor of human genetics at the University of Münster, subsequently establishing what was to become the most important research institute for human genetics in the new German republic. Verschuer's friend and colleague Fritz Lenz, who had run the eugenics department at the Kaiser-Wilhelm Institute since 1933, lost his position for an even shorter period: he was offered a professorship at the renowned University of Göttingen as early as 1946, just a few months after the end of the Nazis' reign of terror.

Verschuer and Lenz were not isolated cases. Most professors of medicine and directors of institutes and clinics during the Third Reich maintained their positions after 1945 or were assigned new jobs shortly afterwards. Both staunch Nazis and 'merely' extreme conservatives had put themselves at the service of the Nazi rulers and helped them to achieve their goals on a far larger scale than other German scientists. Now they could resume their research and teaching in the new Federal Republic of Germany, differently worded, perhaps, but in the same spirit.

For many years, little was known about the 'second careers' of Nazi doctors, even within Germany. This is why *Kontinuität und Neuanfang in der Hochschulmedizin nach 1945*, recently published in German by three professors of medicine at the Philipps-University of Marburg, deserves particular attention.

With the help of the Deutsche Forschungsgemeinschaft, they set out three years



Frozen horror: Nazi doctors at Dachau concentration camp experiment on a political prisoner by immersing him in icy water. The picture was an exhibit at the Nuremberg war crimes trials.

ago to examine the history of the medical faculty in Marburg during the era of National Socialism. This in itself is remarkable because medical faculties, like universities in general, have always refused to analyse their fatal involvement in the Nazi regime.

The three professors soon realized that the end of the Third Reich did not mark a clear turning point for their discipline. Before the results of their research into the 1933–45 period were published, they organized a convention last year on the beginnings of medicine at German universities after 1945. The lectures and findings of this convention are now included in the book.

And indeed, all contributions in the book show just how important it is to examine the two periods together. Whether in Marburg, Münster, Göttingen or at other universities in the three Western-occupied zones, and whether for genetics, neurology or psychiatry, there was no real new beginning in the medical profession anywhere in Germany after 1945. Rather, many paths led directly from the Third Reich by way of Allied occupation to the Federal Republic of Germany. Continuity was initially, and first and foremost, a continuity in terms of individuals, as Verschuer, Lenz and many of their colleagues demonstrate.

But these individuals also guaranteed the continuity of research subjects, textbooks and the syllabuses of medical studies. At least one generation of German medical students,

many of whom have also become professors and researchers, were trained in these circumstances. And former Nazi doctors influenced not only universities but also the beginnings of West German health policy.

How could this smooth transition have happened? The articles in the book suggest several answers to this crucial question. One reason was undoubtedly that doctors were urgently needed to provide medical care in devastated postwar Germany. Many hospital directors, in particular, were able to keep their jobs as a result. Equally decisive was the fact that *Entnazifizierung*—the process of finding members of the Nazi party and other incriminated people—as well as their dismissal by the Allies, was soon carried out only half-heartedly, even at universities.

Perhaps the most important reason, however, was the behaviour and attitudes of universities and medical faculties as institutions. Not only did they support politically incriminated individuals and keep many positions free for professors who had initially been dismissed, but they also obstructed or boycotted academics who had been persecuted or driven into exile by the Nazis. Very few regained a foothold in German universities after 1945.

Closely linked to the attitude of institutions was the personal attitude of most medical professionals in postwar Germany. In a sometimes abstruse way, they either tried to

* *Continuity and New Beginnings in University Medical Schools after 1945*

reinterpret their work in the Third Reich, or hid behind the guise of 'apolitical scientists' who had been abused by the Nazis.

This Marburg convention booklet highlights all these continuities and mechanisms using many examples and drawing on many sources. So it is worthy in two respects. It shows the extent to which the medical profession in the new German republic was influenced by the medical profession in the Third Reich. And it suggests that this was true of most other subjects and faculties at West German universities. □

Marco Finetti is on the editorial staff of Deutsche Universitätszeitung: Das unabhängige Hochschulmagazin, Königswinterer Strasse 418, D-53227 Bonn, Germany.
e-mail: finetti@raabe.bonn.com

Emotional displays

About Face

by Jonathan Cole

MIT Press. Pp. 220. \$25, £14.95

Stuart Sutherland

The case history is rapidly replacing the novel as the favourite reading of the literati. The master of the genre is, of course, Oliver Sacks, who writes with clarity and style, and moves his readers without giving way to sentimentality.

Jonathan Cole believes that facial expressions are the main way to communicate emotion, and to prove his point he has had the ingenious idea of examining the deficits of those who either cannot interpret the expressions of others (the blind and the autistic) or cannot display such expressions themselves (those with Möbius syndrome, Bell's palsy, Parkinson's disease or severe disfigurement of the face caused by burns). Their interactions with others are severely impaired and, feeling like freaks, they may become lonely and desolate.

Some of the comments of the subjects are interesting: for example, blind people who at one time had sight strive to preserve their fading visual images, particularly of those dear to them. Unfortunately, Cole records their remarks indiscriminately: it is hardly news that those blind from birth worry less about their condition than those blinded at a later age.

He goes further than this, however, by arguing speculatively that if we cannot perceive emotions in others we do not learn to express them adequately ourselves. He also thinks that inability to express emotion reduces the emotion felt, a



Vestiges of creation

What is a monster? Is it a thing with a hairy face, extra eyes and missing limbs — or simply anything we do not understand? In *Special Cases: Natural Anomalies and Historical Monsters* (Chronicle, February 1998), Rosamond Purcell tells some macabre legends of monstrosity, ghoulish gossip of scientists and tall tales of early travellers. These plaster life casts of the faces of people from the Indonesian island of Nias (below) were "made by the nineteenth-century Dutch colonial anthropologist J. P. Kleiweg de Zwaan to support his theory of the diversity of evolutionary types within restricted geographical boundaries". The twins joined at the pelvis (right) have "two heads, two upper bodies and a (very rare) third vestigial leg".



belief that is supported by some of his subjects' responses. Further confirmation is provided by Antonio Damasio's work on the importance of bodily reactions in emotion. Indeed, William James's views, scoffed at for years, are now resurgent: "We do not run because we are afraid, we are afraid because we run."

Throughout the book, Cole is anxious to prove that the face is the key to understanding others. At least five facial expressions are universally displayed and universally recognized, and these must be innate. Moreover, although Cole does not mention it, there are single nerve cells in the temporal lobe each of which fires to a specific facial expression regardless of who is making it.

Cole suggests that those disadvantaged with respect to facial expression can be helped — "taught to cope" in psychotherapeutic slang. Indeed, he was partly instrumental in setting up an organization, Changing Faces, to provide such therapy. Certainly, bringing together people with the same affliction is likely to be helpful, if only because it makes

them realize they are not unique. But Cole is vague about other therapeutic methods except for exhortations to those afflicted not to undervalue themselves and encouragement to find something they can do well. He believes that 'denial' — refusing to accept that there is a problem — is a poor strategy, but the point is not well argued and some recent techniques in psychotherapy, admittedly of unproven value, amount to inducing systematic denial.

One of the problems with *About Face* is that, at least in the field of psychology, Cole is an innocent. He asks his subjects many leading questions to which they might accede out of politeness or even bewilderment.

Nor does he understand that people are remarkably bad at attributing causes for their actions. His belief that the face is all-important leads him to suggest that the lack of normal emotions in autism stems from an inability to interpret others' facial expressions rather than, as is usually thought, the other way round. The standard hypothesis is supported by a great deal of evidence that he does

not adequately discuss. As for his case histories, they are sometimes rather inconsequential and he inserts too many comments about himself. Few readers will be fascinated by the news that he had to break off an interview to visit the lavatory.

Case histories can undoubtedly be a source of hypotheses, but to have validity the hypotheses need to be refined and tested by experiment. After all, Freud, relying on case histories, produced the most spectacularly wrong theory of the century; but then, although dealing with emotion, he never looked a patient in the face.

Cole is undoubtedly well-intentioned and has revealed an important and neglected problem. His book is worth reading for that, for the occasional insights his subjects provide and for the interesting but unproved theoretical haes that he raises, even if his prose does not match that of the master, who, incidentally, gives *About Face* his imprimatur. □
Stuart Sutherland is at the Laboratory of Experimental Psychology, University of Sussex, Falmer, Brighton BN1 9QG, UK.

Also of interest

The Psychology of Facial Expression edited by J. A. Russell and J. M. Fernández-Dols. Sixteen chapters offering broad and up-to-date coverage — ethological, neurobehavioural, developmental, dynamic systems and computational — of the role and function of human facial behaviour. Cambridge University Press, £55, \$74.95 (hbk), £19.95, \$29.95 (pbk).

The steeple people

Creating the Cold War University: The Transformation of Stanford

by Rebecca S. Lowen

University of California Press: 1997. Pp. 316. \$45, £35

David Ritson

The cover to this book shows a dramatic and glacial view of the Stanford University campus in California. Stanford has been a leader in changes that, over the past 50 years, have turned the old tranquil ivory towers of academia into the research powerhouses of today. For this reason, Rebecca S. Lowen, a graduate of the Stanford history department, chose Stanford as both a model and a case history to examine the mechanisms that drove this transformation. Today, of course, at the pinnacle of its dazzling rise, Stanford is home to the first daughter, Chelsea Clinton.

Lowen has done a heroic job of wading through old archives, minutes of university committees, papers amassed by former presidents of the university, records of trustee meetings and so on. In surgical detail she has exhumed old policies and exercises of power sometimes benevolent and amazingly far-

sighted and at other times vicious and destructive. Her objective is to understand, in so far as it is possible, today's universities. These are the universities with which we, faculty, students and general public, enjoy a love-hate relationship.

Change has resulted in unparalleled facilities, the brightest students, handsome salaries, light teaching loads and beautiful campuses. These changes have not come without a price. Faculty, even tenured members, are implicitly (or explicitly, as is made clear in the book) driven to publish or perish, and judged by how much money or prestige they add to the university, whereas the old humanistic disciplines are treated as poor cousins of engineering, computer science, law, medicine and the physical sciences.

The simplistic explanation is that the new patrons, the military-industrial complex, and a new breed of entrepreneurial academic scholars combined to produce this out-come. Lowen's underlying and well-documented thesis is to detail the extraordinary, and perhaps dominant, role of university administrations in this process.

The book starts in the 1930s and follows through to the late 1960s. Major change came in the wake of the Second World War. The university president in those days was the jovial and friendly J. Wallace Sterling. For most of these years he was teamed with the tough, far-sighted and sometimes abrasive provost Frederick Terman.

Together, Sterling and Terman successfully played 'fat man/thin man' roles. During the Second World War, Terman had directed the Harvard Radio Research Laboratory. After his return to Stanford he became dean of the school of engineering. From 1956 to the mid-1960s he was provost. He developed and labelled the strategy "steeples of excellence".

This was aimed at steering the university away from trying to be good at everything. Instead the goal was to exploit government and industrial sources for funding and to channel faculty and university resources into a few areas where the university could become a leader. Terman's policies also produced the strong synergy between industry and Stanford that spawned the rise of today's Silicon Valley. The methods used to accomplish these aims are detailed by Lowen in a chapter entitled, not unexpectedly, "Building steeples of excellence".

Lowen has clearly tried hard to keep the book objective, and she generally succeeds. As a 'modern' historian she primarily relies on the written contemporaneous records. These have their limitations. Nobody, even in private letters and memos, is going to admit to anything but the loftiest motives in pursuing their aims. She, as do we all, obviously has a nostalgic hankering for a Camelot when scholars really lived in ivory towers and thought great thoughts. In this sense she has

New in paperback

Camping With the Prince and Other Tales of Science in Africa

by Thomas A. Bass

Moyer Bell, \$13.95

"Bass's stories highlight the often disastrous effects of ignoring local knowledge and practices", Len Goodwin, *Nature* 344, 901 (1990)

Life's Grandeur: The Spread of Excellence from Plato to Darwin

by Stephen Jay Gould

Vintage, £7.99

Published in the United States as *Full House*.

"Gould presents his ideological views in characteristic evangelical style", Christian de Duve, *Nature* 383, 771 (1996)

Yellow Fever, Black Goddess: The Coevolution of People and Plagues

by Christopher Wills

Helix, \$13

"To the point, scholarly, disinterested and rich in technical detail... provides pleasure and stimulation", V. Geist, *Nature* 386, 236 (1997)

Dinosaurs of the Flaming Cliffs

by Michael Novacek

Anchor, \$14.95

"A thoroughly absorbing adventure story.... This eloquent and accessible blend of science and discovery should appeal to a wide readership... and may well inspire the next generation of palaeontologists", Angela Milner, *Nature* 384, 426 (1996)

The Logic of Failure: Why Things Go Wrong and What We Can Do to Make Them Right

by Dietrich Dörner

Addison-Wesley, \$15

"A lively treatise, accessible to the lay reader [which] capitalizes on real-life cases and refined *ad hoc* experiments", M. Piattelli-Palmarini, *Nature* 382, 505 (1996)

One River: Explorations and Discoveries in the Amazon Rain Forest

by Wade Davis

Simon and Schuster, \$16

"The author has woven a biographical tapestry rich in history, adventure, intrigue and scholarship... an engaging and provocative read", Michael J. Balick, *Nature* 384, 229 (1996)

The Clock of Ages

by John J. Medina

Cambridge University Press, £9.95, \$15.95

"A tour of human ageing that aims to educate and entertain.... The science is adequate but too often fails to do justice to this complex and challenging topic", Tom Kirkwood, *Nature* 382, 506 (1996)

been, perhaps, less than fair to university administrations who have a duty to encourage and open up new directions in scholarship.

Disappointingly, the book stops before the start of the Vietnam war and the 1960s' 'student revolt'. It also fails adequately to emphasize the role of university 'overhead' charges on contracts, an important financial support mechanism. The last chapter on undergraduate teaching is the weakest. Lowen has documented well the then administration's disregard for undergraduate teaching, but her treatment is confined to a short chapter and seems to be coloured by her own Stanford experiences.

Who should read this book? Certainly Stanford faculty members, if only to satisfy their voyeurism. Certainly all university faculty and administrators who wish to understand the nuts and bolts of wielding power. More generally, it is a provocative and literate book of interest to anybody who wants to understand the strengths and weaknesses of our modern higher-education systems. □

David Ritson is in the Department of Physics, Stanford University, Stanford, California 94305-4060, USA.

ADVERTISEMENT

"Beside good science, there is much humour and much of interest outside the field of science. ... In my case, it kept me awake and delighted until the small hours"

Keith J. Laidler, Professor Emeritus of Chemistry, Ottawa University

The Bedside Book Genius and Eccentricity in Science 1869–1953

Edited by Walter Gratzner, with a foreword by Stephen Jay Gould

A fascinating selection of articles from *Nature*, offering a panorama of science seen against the background of 19th and 20th century history.



Price: £19.95/US\$29.95 + postage
Discount rate available for current subscribers.
To order, or for more information,
contact: Nature Marketing
Tel. +44 (0) 171 843 4985
Fax. +44 (0) 171 843 4998
E-mail: subscriptions@nature.com

In retrospect chosen by Kevin Padian

Tess of the D'Urbervilles

by Thomas Hardy
(1891)

Thomas Hardy was 19 when Charles Darwin published *On the Origin of Species* in 1859, and it profoundly affected his view of the relationship of humans to nature. He endured criticism throughout his life for his fatalistic plots and the tragic ends of his characters. But these devices didn't reflect Darwin's influence. Like Darwin, Hardy saw nature as morally neutral, despite its constant destruction of life. Only humans, with their heartless social systems, can show malice and hurt others wantonly. And as Darwin taught us, sexual selection, more than natural selection, drives human evolution.

Hardy recognized intuitively that Darwin's *magnum opus* was not only about the population processes of heredity and selection, but about evolutionary legacy and deep time. From the opening page of *Tess* to its closing paragraph, evolution is the breath of the novel.

It begins as Tess's father, the labourer, John Durbeyfield, is informed by the parson, an amateur genealogist, that he is descended from noble blood (the ancient D'Urbervilles). Tess is sent to the local manor to see her 'cousins' in the hope of a handout, but the true D'Urbervilles are long extinct; the ancient title has been purchased by a parvenu tycoon, and Tess's false cousin ultimately rapes her.

As the novel develops, we find that Tess's perseverance comes not only from her peasant strength but also from an atavistic ferocity inherited from her remote ancestors, knavish, pugnacious aristocrats who mistreated the local girls in their own times. By contrast, Tess's false 'cousin' Alec fails at everything in life partly because he has no genetic reserves to draw upon.

Evolutionary scales in *Tess* run from the genetic to the geological. Fertility, the basis of genetic change, permeates the novel; it is, after all, the business of rural communities. The fields and glens that Hardy describes are always buzzing with bees and butterflies, exploding with pollen- and nectar-laced flowers, vitalized by rains and streams:

"The season developed and matured. Another year's instalment of flowers, leaves, nightingales, thrushes, finches, and such ephemeral creatures took up their positions where only a year ago others had stood in their place when these were nothing more than germs and inorganic particles. Rays from the sunrise drew forth the buds and stretched them into long stalks, lifted up sap in noiseless streams, opened petals, and sucked out scents in invisible jets and breathings."

Traditions of fertility also extend to cultural legacy. When we first meet Tess, she is participating in the annual May Day walk of the last women's club in England to preserve the tradition (note the harbinger of extinction). The women wear virginal white gowns, but the traditions, like the whiteness of their dresses, have degenerated to the point that



aged matrons long past their fertile years walk with the young girls. They carry a willow wand in one hand and flowers in the other, dual symbols of male and female fertility on a day celebrating fertility itself. But the women have no

consciousness of the irony of these symbols because their associations have been forgotten.

Time — in the sense of John MacPhee's 'deep time', numberless, incomprehensible eons — shapes everything in the present, at one scale or another. Hardy is constantly moving among these scales, here pointing out a Gothic church, there reminding us of the ancient age of the stones used to build it.

Hardy often begins his stories with a road, along which travellers are making their way. But before the story proceeds, the road is transformed into a Roman way, or a far more ancient path or track, or it passes a pre-Christian burial mound, or cuts through Palaeozoic strata that impassively regard the traveller's progress. These forms are ancient, traditional and beyond human memory. Hardy reminds us that we are individually ephemeral; yet the actions of groups of humans, like geological processes, gradually shape the landscape, construct farms, fields, towns, cultures, laws, history itself. It is as if Hardy were tending a cosmic ant-farm.

Tess, however, is no ordinary ant, but a heroine of grand and tragic dimensions rooted in evolutionary constraint and genetic possibility. Her ancestors are buried like fossils in Kingsbere-sub-Greenhill (King's-bier-below-green-hill), as fitting a locality for a crypt as one could ask. Yet Tess carries the D'Urberville legacy in her flashing temper, her determination and her unwavering devotion to a middle-class husband.

In the end, Tess mortally wounds her false cousin Alec for his deceits, and flees with her husband Angel across the moor until they wind up, exhausted, at Stonehenge (where else?). The ancient geological sarsens stand over her like sentinels, as she sleeps like a Druid priestess on an altar-like stone until the police arrive.

The novel closes as an offstage bell tolls notice of Tess's hanging. There is one final evolutionary irony as Angel walks off with Tess's sister, who of course carries more genetic similarity to Tess than anyone in the world. Victorian laws would have prevented their union; but if we know one thing about Hardy's characters, it is that they seldom applied the wisdom of the animal husbandry they knew so well to their own breeding habits.

Kevin Padian is in the Department of Integrative Biology, University of California, Berkeley, California, 94720-3140, USA.

Real-time seismology and earthquake hazard mitigation

Hiroo Kanamori, Egill Hauksson & Thomas Heaton

Recent advances in seismic sensor technology, data acquisition systems, digital communications, and computer hardware and software make it possible to build reliable real-time earthquake information systems. Such systems provide a means for modern urban regions to cope effectively with the aftermath of major earthquakes and, in some cases, they may even provide warning, seconds before the arrival of seismic waves. In the long term these systems also provide basic data for mitigation strategies such as improved building codes.

If the impact of future earthquakes is to be effectively reduced, seismic hazards need to be addressed on several different timescales. On the timescale of decades, land use regulations and building codes need to be improved. On the timescale of a few years, measures aimed at preparing for possible earthquakes should be encouraged at personal and community levels. On shorter timescales, months to days, accurate earthquake predictions of size, location and time would be required. However, because of the extreme complexity involved in earthquake processes, reliable short-term prediction is not possible at present (see Box 1). Furthermore, even if such prediction should become possible, large earthquakes in densely populated urban areas are likely to cause extensive damage and disruption to society. To minimize the immediate impact of large earthquakes in such areas, we advocate a strategy of taking full advantage of recent advances in seismology, sensor, computer and telemetry technologies for developing rapid and reliable real-time

earthquake information systems. Despite improvements in engineering designs of individual structures, modern urban and suburban areas as a whole are more vulnerable to earthquakes than ever.

The purpose of these modern earthquake information systems is to give rapid notification of the earthquake parameters (time, location and magnitude) and estimates of ground motion (acceleration, velocity, displacement, spectral amplitude and so on), in order to assess where emergency response is most needed and to estimate the overall societal impact of the event. These systems will also contribute to shortening the recovery time following major earthquakes by allowing rapid determination of what resources are needed for rebuilding and recovery after the event¹⁻³. In highly industrialized areas, minimizing the response time after an earthquake is important for search and rescue efforts aiming to protect life and property. Further reducing the recovery time is critically important for utilities, communication, transportation, financial companies, commerce and local industries. The recent earthquakes in Northridge, California, and Kobe, Japan, clearly demonstrated the need for such systems, where both the response time and recovery time could have been shortened had sufficient information been available. For example, during the Kobe earthquake, because seismic networks and communication systems were disrupted by strong shaking, the full extent of the damage was not known to the central government in Tokyo until many hours later.

Today, real-time systems are capable of providing basic earthquake information within minutes, and in the future this could be reduced to tens of seconds. In some cases, facilities will receive this information even before ground shaking begins (early warning). This may allow for clean emergency shutdown or other protection of systems susceptible to damage, such as power stations, transport and computer systems. As an example of the use of early warning technology, the telecommunication networks are interested in collecting data on network performance during and immediately following a major earthquake. Such data are needed for designing future improvements of the telephone systems.

Perhaps the strongest rationale for deploying an extensive network of versatile seismic stations is that it will lead to a much better understanding of earthquake ground shaking. Ground motions from a variety of earthquake sizes and locations will allow us to distinguish the effects of wave propagation (which do not vary for earthquakes in the same location) from the effects of the earthquake rupture. This will ultimately provide much better predictions of the distribution of ground motions in future earthquakes. Furthermore, quantitative measures of building performance in earthquakes can only be achieved if shaking during damaging earthquakes is recorded with spatial density sufficient to characterize the ground motion.

Box 1 Earthquake prediction

In a narrow sense, an earthquake is a sudden failure process; in a broad sense, it is a long-term, complex process of stress accumulation and release occurring in a highly heterogeneous medium, the Earth. Considerable advances have been made in the past several decades in understanding crustal deformation and stress accumulation processes due to plate motion, rupture dynamics, friction, interaction between faults, fault-zone structures and nonlinear dynamics. Thus, it is possible to predict to some extent the seismic behaviour of the crust in the future from various measurements taken in the past and at present. But even a simple mechanical model of earthquakes exhibits very complex behaviour, suggesting that earthquakes are essentially a chaotic process and are predictable only in a statistical sense²². In many conceptual models, the magnitude of an earthquake is determined by the length of the rupture. In these models, every small earthquake that occurs has the potential to grow into a large one, but in most cases the rupture arrests in a short distance²³. As there are 100,000 times more earthquakes of magnitude 2 (M2) than of magnitude 7, it may not be possible to predict which M2 will grow into an M7 earthquake.

Even if the physics of earthquakes were well understood, the obvious difficulty of making detailed measurements of various field variables (structure, strain and so on) in three dimensions in the Earth would still make accurate deterministic predictions unlikely. Thus, a short-term prediction, if any is made, is bound to be very uncertain. Such uncertain predictions might be useful for those places where the social and economical environments are relatively simple, and false alarms and missed events can be socially tolerated. However, in modern highly industrialized urban areas with complex lifelines, communication systems and financial networks, such uncertain predictions might not be useful; they could inadvertently damage local and global economies.

Real-time earthquake information system

The idea of using rapid earthquake information for emergency

operations is not new^{4,5}, and several systems have been developed in Japan, Mexico, Taiwan and the United States. The following are some examples.

Japan. In the late 1950s, simple seismometers were installed for a railway alarm system. Since the operation of the Bullet Train started in 1964, an automatic system to stop or slow down trains during strong earthquakes has been developed. The most recent system, UrEDAS, uses a sophisticated algorithm for seismic detection and location, and has been developed for use by the Japanese railway system^{6,7}. Some utility companies have also developed real-time ground-motion detection systems for emergency services for their own facilities⁸. Recently the City of Yokohama embarked on a project to deploy a 150-station real-time strong-motion network⁹.

Mexico. In 1985, an earthquake of magnitude $M = 8.1$ in the Michoacan seismic gap, about 320 km west of Mexico City, caused very heavy damage to the city. As a similar large earthquake is expected in the Guerrero seismic gap, about 300 km southwest of Mexico City, a seismic alert system (SAS) was developed in 1991 with funding from a private foundation¹⁰. This system has a specific objective: to detect $M > 6$ earthquakes in the Guerrero gap with a seismic network deployed in the gap area, and issue an early warning of strong ground motion to the residents and authorities in Mexico City. As it takes about 100 seconds for seismic waves to travel from the Guerrero area to Mexico City, this system could provide an early

warning with up to 60 seconds lead time. When a $M = 7.3$ earthquake occurred on 14 September 1995 in the Guerrero gap, this system successfully broadcast an alarm on commercial radio stations in Mexico City about 72 seconds before the arrival of strong ground motion.

Taiwan. Two prototype earthquake early warning systems have been set up in Taiwan, one for a local area near Hualien, and another for the entire island^{11–13}. These systems use state-of-the-art seismic network technology, and are designed to provide critical information on earthquakes and resulting ground motions for various emergency and recovery operations.

United States. Rapid notification systems had been in internal use by the US Geological Survey (USGS) in California since the early 1980s. In 1990, the California Institute of Technology (Caltech) and the USGS Pasadena Office initiated the CUBE project (CUBE stands for Caltech/USGS Broadcast of Earthquakes) (Fig. 1)¹⁴. It was realized that closely coordinated efforts between academia, governments and private companies are essential for effective earthquake mitigation in modern metropolitan areas such as Los Angeles. One of the important objectives of the CUBE project was therefore to promote user education and close coordination efforts between various organizations with a rapid and reliable earthquake information system during major earthquake sequences. In 1993, the University of California at Berkeley and the US Geological Survey Office in Menlo Park developed a system known as REDI (the Rapid Earthquake Data Integration project) to broadcast earthquake data in northern California¹⁵.

Real-time information can be used in many different ways. The following are some examples. (1) During the aftershock sequence of the 1989 earthquake in Loma Prieta, California ($M = 6.9$), the US Geological Survey implemented a warning system to protect construction workers cleaning up the collapsed freeways in Oakland, about 100 km from the epicentre¹⁶. When large aftershocks occurred, this system provided about 20 seconds of warning to the workers so they could evacuate from the potentially hazardous area. (2) During the 1991 Sierra Madre earthquake near Pasadena, workers on the Santa Fe Railway, upon receipt of CUBE information, were able to reduce the routine inspection time by 2 hours. (3) Utility companies usually do not have a large field staff for emergency operations for relatively infrequent events such as earthquakes. In an emergency, therefore, it is important for them to decide quickly on allocation of limited resources on the basis of the magnitude of the event and distribution of ground-motion intensity. (4) Some companies use earthquake information to locate the sites of possible leakage of hazardous materials and broken pipes. (5) During the 1994 Northridge, California, earthquake, a large transformer of a power company was seriously damaged by one of the larger aftershocks. CUBE helped the engineers to respond quickly enough to avoid the loss of a severely damaged transformer, thereby saving a few million dollars. (6) Immediately after a large earthquake, local telephone service, even if it is physically intact, is often disrupted because of cascading failures caused by a sudden increase in telephone traffic. If real-time data on strong ground motion is available, telephone companies may be able to avoid this problem by isolating the overloaded parts of their network. These are just a few examples, and when real-time information becomes more widely available, more applications are sure to be found.

In the following, we discuss some technical issues using the CUBE system as an example; these issues are common to many of the existing real-time systems mentioned above. CUBE uses the information provided by a 250-station seismographic network, the Southern California Seismic Network (SCSN), and more than 50 advanced digital seismographic stations, including those from TERRAScope, a state-of-the-art broad-band digital seismic network with wide dynamic range. These networks are jointly operated by Caltech and USGS. To set up the CUBE project, some hardware and software modules were added to the existing system so that earth-

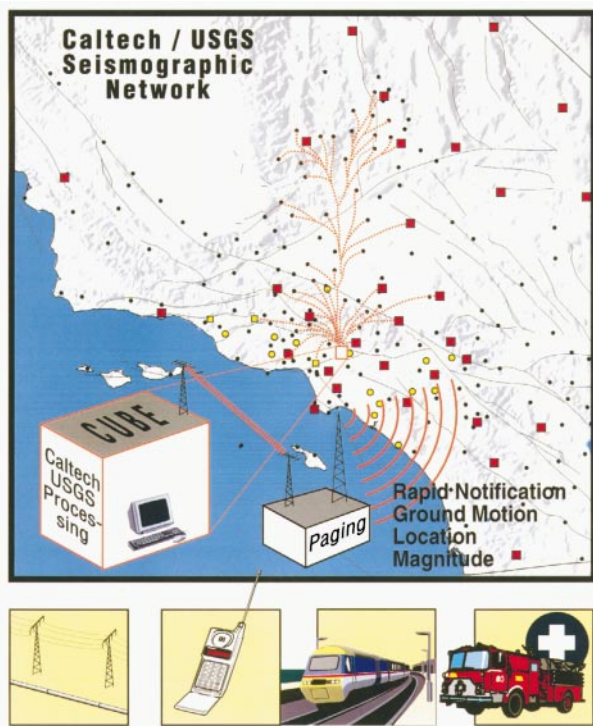


Figure 1 Schematic diagram of the Southern California Seismographic Network (SCSN), a joint project of the US Geological Survey and the California Institute of Technology, and CUBE. The signals from the analog field seismic stations (those of the SCSN are shown as small dots, and digital stations such as TERRAScope by red squares, yellow circles and squares) are transmitted in real time to the central processing facility at Caltech in Pasadena through analog or digital telephone lines, radios, and microwave links (indicated by dashed red lines). The data are automatically processed at Pasadena, and when an earthquake is detected, its location and magnitude are determined within a few minutes. The earthquake parameters and peak ground-motion values are sent to project participants by means of a commercial paging system. The boxes at the bottom illustrate the diverse categories of the users. The pagers can also be connected to personal computers to display the location or the amplitude of ground motions on a map showing highways, railways and other infrastructure facilities.

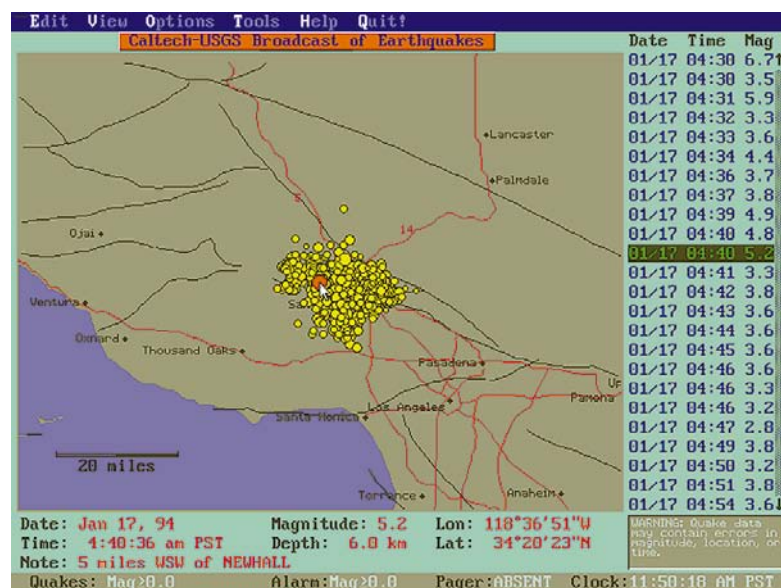


Figure 2 The aftershock sequence of the 1994 Northridge, California, earthquake ($M = 6.7$) shown on the display that CUBE users see. The origin times of the events are listed on the right, and information on the hypocentre of the selected event (red) is displayed at the bottom. Although reporting of the Northridge mainshock was delayed (see text), the display of the aftershock activity shown here helped the emergency operators to respond quickly both to the mainshock and the large aftershocks.

quake information, such as the epicentre (location), magnitude and origin time, could be broadcast to project participants over a commercial paging system (Figs 1 and 2). Not only large earthquakes, but also all earthquakes above a preset threshold magnitude are reported, to produce a continuous display of seismicity.

In the beginning of the CUBE project, four organizations (Santa Fe Railway, Metropolitan Water District, City of Los Angeles Department of Water and Power, and Southern California Edison) participated in the project and installed the CUBE display system in their facilities' operation room. CUBE now has about 30 participating organizations and is generally considered as an integral part of the emergency preparedness plan for southern California. One of the critical elements of this project is close interaction between the CUBE operator (Caltech/USGS) and the project participants. More than 100 participants attend CUBE user meetings which are held at Caltech twice a year, and the feedback from the users plays an important role in improving the system.

The CUBE system was used effectively during the 1991 Sierra Madre earthquake ($M = 5.8$), the 1992 Joshua–Tree earthquake ($M = 6.4$), the 1992 Landers–Big Bear earthquake ($M = 7.3$) and the 1994 Northridge earthquake ($M = 6.7$). During the Northridge earthquake, failure of a key telephone line and overloading of the system caused delay in broadcasting the information on the main shock, but the spatial and temporal distributions of the ensuing aftershock activity were used by the participating agencies for guiding various emergency operations. This illustrates the importance of sending information for not only large earthquakes but also the background seismicity and aftershocks. A continuous display system is also useful for educating the users about the regional seismicity and tectonics. As large events are rare, a sudden alarm for a large event by itself would be difficult to deal with quickly unless the users were trained and could evaluate the new event location in relation to the background seismicity.

Since 1994, CUBE and REDI have joined forces to broadcast earthquake information for the entire state of California. This joint broadcast system, often called CUBE/REDI, is the only system that involves many different types of users distributed across so large an area.

New technology and methods

In the past decade significant advances have been made in seismic sensors, digitizers, communication systems, computers and seismological methods. Introduction of these new technologies and methods can greatly enhance the capability of real-time systems such as CUBE/REDI.

As part of TERRAScope and other projects, a combination of modern broad-band seismometers, accelerographs and 24-bit digital data loggers is being used to replace the short-period sensors of the SCSN. These systems can record motion ranging from the very weak, less than $10^{-9}g$, to very strong, for instance $2g$ (where g is the acceleration due to gravity at the Earth's surface). The broad-band capability is important for understanding the nature of earthquakes and resulting ground motions, which eventually will provide a better estimation of strong ground motions from future very large earthquakes (Box 2).

High-capacity digital communication systems, both land lines from the local telephone service providers (such as Pacific Bell/GTE Frame Relay) and spread spectrum radios, that are capable of sending large amounts of data rapidly and reliably are now available (Fig. 1). As the required data rate for a typical modern digital seismic station is about 20 kbps (kilobits per second), these modern

Box 2 Why broad-band seismographs?

Two types of seismographs are used in modern seismic networks. The first is the digital strong-motion accelerograph which records mainly strong ground motions exceeding 1% of g . Although the frequency band is limited, usually higher than 0.1 Hz, it provides useful data for most engineering applications. The second is the broad-band seismograph which covers a frequency band from 0.001 Hz (a period of 1,000 s) to 10 Hz, or even wider. Unfortunately, these instruments are driven off-scale when the ground motion velocity exceeds 1 cm s^{-1} . As the strong ground motion can exceed 1 m s^{-1} , these instruments are not adequate for direct strong-motion applications. Nevertheless, high-quality ground-motion data from small to moderate earthquakes recorded with broad-band instruments are important for estimating strong ground motion from very large earthquakes in the future. Because ground-motion data recorded in the proximity of large earthquakes are scarce, our knowledge about such ground motion is very limited; yet demand for such knowledge is increasing rapidly as many large structures such as high-rise buildings, bridges and large storage tanks are being built. Seismologists can now construct a model for a large earthquake as proper superposition of small to moderate earthquakes and quantitatively estimate ground motion from very large earthquakes using the data from small to moderate earthquakes. For this reason, deployment of broad-band seismographs and archiving of their data are crucial.

communication systems are more than adequate for transporting data for real-time seismic systems.

Another important consideration pertains to whether data telemetry should be continuous or event-triggered. In 'triggered' systems, only the portion of data that contains earthquake ground motion is telemetered from a field station to the central processing facility. This minimizes the telemetry cost, but the work-load on the system increases suddenly during a major earthquake, and could cause a system failure during the time when real-time data are most needed for emergency services. In contrast, in systems using continuous data transmission to the central facility, data are processed regardless of whether earthquakes are occurring. This minimizes the fluctuation of work-load, leads to more reliable event identification using the data from all the stations simultaneously, and assures more reliable operation of the system. Also, continuous data are easier to archive than triggered time series which can be fragmented and have unexpected transients at the beginning of a time window of data. Thus, if the capacity and cost of telemetry system allow continuous telemetry, it is preferable to the triggered mode.

Modern computers with multiple central processing units provide enough capability for processing data for real-time seismic systems. Telemetry between the stations and the central processing facility can be viewed as inter-computer networking which allows two-way communication to aid in diagnosis, maintenance and calibration of stations. A similar digital telemetry can be used to send information from the central facility to remote user facilities.

Data processing methods need to be expanded to accommodate users' needs. For example, for immediate emergency relief after a earthquake, it is often more important to know the spatial distribution of strong-motion parameters such as peak acceleration, peak velocity and spectral amplitudes.

Traditionally, seismic networks provide the location and the origin time of earthquakes determined from the arrival times of various seismic phases, such as P and S phases. In this case 'the location' refers to the point where the rupture initiates (the hypocentre). This practice will continue to be central to any real-time system. However, for very large earthquakes, the energy is radiated from a large area surrounding the entire fault plane, and the hypocentre indicates only where rupture starts. Thus, it is desirable to locate, in addition to the traditional hypocentre, the centre of the energy radiation, here referred to as the ground-motion centroid. In essence, the centroid must be close to the station with the largest amplitude. In practice this method has long been used by seismologists who guessed where the earthquake was from the reports of shaking intensity. With modern seismic instruments which can measure ground motion accurately, the method can be computerized to locate the strong-motion centroid¹⁷.

Future development

Real-time systems must be reliable and free from false alarms. As the speed and reliability of information delivery increase, these systems will be capable of issuing rapid notifications, ground motion maps and early warning to some users. Recording and archiving broadband data from small to large earthquakes is also important for estimating ground motion from very large earthquakes, which in turn is crucial for improving engineering practice, especially for large buildings and structures. Real-time systems with modern data archiving capability can also be used to accomplish this longer-term goal. Incorporating these concepts, Caltech, USGS and the California Division of Mines and Geology recently initiated a collaborative project, TriNet, with the goal of building a state-of-the-art real-time earthquake information system in southern California¹⁸.

Another important element is the development of tools for modelling damage to buildings and lifelines, and estimating casualties in near real-time, given the source parameters and ground-motion maps of an earthquake. Software based on GIS (the

Geographical Information System) is currently being developed for this purpose. For example, the California Office of Emergency Services (OES) helped to secure 8.6 billion dollars in federal disaster relief assistance based on an estimated total loss of 15 billion dollars for the Northridge earthquake. This estimate was produced within one day of the earthquake and was based on modelled ground shaking and databases of building inventories¹⁹. Development of real-time damage assessment tools has continued since the Northridge earthquake, and the California OES currently operates a GIS-based system that will provide damage projections within several hours of the occurrence of an earthquake²⁰. Furthermore, the GIS systems provide the necessary platform for including multi-hazard information from other types of events such as fire, flood and wind²¹.

Overall, the introduction of modern broad-band seismometers, digital data loggers, communication systems and computers holds promise for building rapid and reliable systems for real-time monitoring of the distribution of strong ground motion, the essential for emergency operations and early warning systems. However, effective mitigation of earthquake hazards cannot be achieved by science and technology alone. Mitigation efforts should be closely coordinated between various emergency organizations and utility, transport and communication companies. Rapid, reliable earthquake information systems will play a key role in organizing such efforts and greatly contribute to reducing the hazard through more rapid and focused emergency response in modern urban areas. □

Hiroo Kanamori, Egill Hauksson and Thomas Heaton are at the Seismological Laboratory, California Institute of Technology, Pasadena, California 91125, USA.

1. National Research Council *Real-time Earthquake Monitoring, Early Warning and Rapid Response* 1–52 (National Academy Press, Washington, DC, 1991).
2. Ellsworth, W. & Heaton, H. Real-time analysis of earthquakes: Early-warning systems and rapid damage assessment. *Sensors* (April) 27–33 (1994).
3. Noda, S. & Meguro, K. A new horizon for sophisticated real-time earthquake engineering. *J. Natural Disaster Sci.* 17, 13–46 (1995).
4. Heaton, T. H. A model for a seismic computerized alert network. *Science* 228, 987–990 (1985).
5. Nakamura, Y. & Tucker, B. E. Japan's earthquake warning system: Should it be imported to California? *California Geol.* 41, 33–41 (1988).
6. Nakamura, Y. On the Urgent Earthquake Detection and Alarm System (UrEDAS). in *Proc. 9th World Conf. Earthquake Eng.* VII, 673–678 (Jpn Assoc. for Earthquake Disaster Prevention, Tokyo, Kyoto, 1988).
7. Nakamura, Y. Real-time information systems for hazards mitigation. in *Proc. 11th World Conf. Earthquake Eng.* CD-ROM, Pap. No. 2134 (Pergamon, Oxford, 1996).
8. Cret, A. & Katayama, T. On-line damage estimation for lifeline systems during earthquakes. in *Proc. 10th World Conf. Earthquake Eng.* 9, 5501–5504 (Balkema, Rotterdam, 1992).
9. Kikuchi, M. High-density strong-motion network in a large city (in Japanese). *Gakaku* 66, 841–844 (1996).
10. Espinosa-Aranda, J. M. et al. Mexico City Seismic Alert System. *Seismol. Res. Lett.* 66, 42–53 (1995).
11. Shin, T. C., Tsai, Y. B. & Wu, Y. M. Rapid response of large earthquakes in Taiwan using a real-time telemetered network of digital accelerographs. in *Proc. 11th World Conf. Earthquake Eng.* CD-ROM, Pap. No. 2137 (Pergamon, Acapulco, 1996).
12. Lee, W. H. K., Shin, T. C. & Teng, T. L. Design and implementation of earthquake early warning systems in Taiwan. in *Proc. 11th World Conf. Earthquake Eng.* CD-ROM, Pap. No. 2133 (Pergamon, Acapulco, 1996).
13. Teng, T.-L., Wu, L., Shin, T.-C., Tsai, Y.-B. & Lee, W. H. K. One minute after: Intensity map, epicenter, and magnitude. *Bull. Seismol. Soc. Am.* 87, 1209–1219 (1997).
14. Kanamori, H., Hauksson, E. & Heaton, T. TERRAScope and CUBE project at Caltech. *EOS* 72, 564 (1991).
15. Gee, L. S. et al. Real-time seismology at UC Berkeley: The rapid earthquake data integration project. *Bull. Seismol. Soc. Am.* 86, 936–945 (1996).
16. Bakun, W. H., Fischer, E. G., Jensen, E. G. & VanSchaack, J. Early warning system for aftershocks. *Bull. Seismol. Soc. Am.* 84, 359–365 (1994).
17. Kanamori, H. Locating earthquakes with amplitude: application to real-time seismology. *Bull. Seismol. Soc. Am.* 83, 264–268 (1993).
18. Mori, J. et al. TriNet: State-of-the-art seismic and strong-motion network in southern California. *EOS* (submitted).
19. Goltz, J. D. Use of loss estimates by government agencies in the Northridge earthquake for response and recovery. *Earthquake Spectra* 12, 441–455 (1996).
20. Eguchi, R. T. et al. Real-time loss estimation as an emergency response decision support system. The Early Post-Earthquake Damage Assessment Tool (EPEDAT). *Earthquake Spectra* (in the press).
21. Yamazaki, F., Noda, S. & Meguro, K. Development of earthquake monitoring and early damage assessment systems in Japan. *Bull. Earthquake Resistant Structure Research Center* 30, 45–58 (Institute of Industrial Science, University of Tokyo, 1997).
22. Turcotte, D. L. *Fractals and Chaos in Geology and Geophysics* 1–221 (Cambridge University Press, Cambridge, 1992).
23. Brune, J. N. Implications of earthquake triggering and rupture propagation for earthquake prediction based on premonitory phenomena. *J. Geophys. Res.* 84, 2195–2198 (1979).

Correspondence should be addressed to H.K.

TGF- β signalling from cell membrane to nucleus through SMAD proteins

Carl-Henrik Heldin, Kohei Miyazono & Peter ten Dijke

The recent identification of the SMAD family of signal transducer proteins has unravelled the mechanisms by which transforming growth factor- β (TGF- β) signals from the cell membrane to the nucleus. Pathway-restricted SMADs are phosphorylated by specific cell-surface receptors that have serine/threonine kinase activity, then they oligomerize with the common mediator Smad4 and translocate to the nucleus where they direct transcription to effect the cell's response to TGF- β . Inhibitory SMADs have been identified that block the activation of these pathway-restricted SMADs.

TGF- β 1 is the prototype of a large family of cytokines that includes the TGF- β s, activins, inhibins, bone morphogenetic proteins (BMPs) and Müllerian-inhibiting substance (reviewed in ref. 1) (Table 1). Members of the TGF- β family exert a wide range of biological effects on a large variety of cell types, for example they regulate cell growth, differentiation, matrix production and apoptosis. Many of them have important functions during embryonal development in pattern formation and tissue specification; in the adult they are involved in processes such as tissue repair and modulation of the immune system.

Here we discuss recent breakthroughs in our understanding of the mechanisms used by members of the TGF- β family to elicit their effects on target cells, focusing on the pivotal role of SMAD proteins in relaying signals from cell-surface receptors to the nucleus.

Signalling through receptor complexes

TGF- β family members initiate their cellular action by binding to receptors with intrinsic serine/threonine kinase activity. This receptor family consists of two subfamilies, type I and type II receptors, which are structurally similar, with small cysteine-rich extracellular regions and intracellular parts consisting mainly of the kinase domains. Type I receptors, but not type II receptors, have a region rich in glycine and serine residues (GS domain) in the juxtamem-

brane domain. Each member of the TGF- β superfamily binds to a characteristic combination of type I and type II receptors (Table 1), both of which are needed for signalling.

Studies of the receptors for TGF- β have provided a model for the activation of these serine/threonine kinase receptor complexes². TGF- β 1 first binds to the type II receptor (T β R-II), which occurs in the cell membrane in an oligomeric form with activated kinase^{3,4}. Then, the TGF- β type I receptor (T β R-I), which may also occur in an oligomeric form⁵ and cannot bind TGF- β in the absence of T β R-II, is recruited into the complex; T β R-II phosphorylates T β R-I in the GS domain to activate it. The assembly of the receptor complex is triggered by ligand binding, but the complex is also stabilized by direct interaction between the cytoplasmic parts of the receptors⁶. The model² predicts that the type II and type I receptors act in sequence, which is supported by the finding that a constitutively active T β R-I (Thr204 replaced with an aspartate residue) is able to exert TGF- β signals in the absence of T β R-II (ref. 7). It is likely that other serine/threonine kinase receptor complexes are also activated by a similar mechanism^{8,9}, although some variations on the theme have been noted. One of the TGF- β isoforms (TGF- β 2) binds only with low affinity to T β R-II and requires the cooperation with T β R-I or betaglycan, an accessory transmembrane proteoglycan, for high-affinity binding¹⁰. Moreover, BMPs bind

Table 1 TGF- β family members, their receptors and signalling molecules

Subfamily	TGF- β	Activin	BMP
Examples of ligands	TGF- β 1 TGF- β 2 TGF- β 3	Activin A	BMP-2 BMP-4 BMP-7/OP-1
Type II receptors	T β R-II	ActR-II ActR-IIB	BMPR-II ActR-II ActR-IIB
Type I receptors	T β R-I	ActR-I? ActR-IB	BMPR-IA BMPR-IB ActR-I
Pathway-restricted SMADs	Smad2 Smad3	Smad2 Smad3	Smad1 Smad5 Smad9?
Common-partner SMAD	Smad4	Smad4	Smad4
Inhibitory SMADs	Smad6 Smad7	Smad6 Smad7	Smad6 Smad7
Responses	Inhibition of mitogenicity Induction of extracellular matrix	Induction of dorsal mesoderm Induction of erythroid differentiation Induction of follicle-stimulating hormone release	Induction of ventral mesoderm Induction of cartilage and bone Induction of apoptosis

The three best-characterized vertebrate TGF- β subfamilies are listed.

with low affinity to BMP type I or type II receptors individually, and with high affinity only when the two BMP receptor types are presented together^{11–13}.

Analysis of ¹²⁵I-labelled TGF- β 1 crosslinked to its receptors has suggested that the signalling complex is a heterotetramer consisting of two T β R-I and two T β R-II molecules¹⁴. This conclusion is supported by experiments using chimaeric erythropoietin/TGF- β receptors which showed that both homodimerization of type I receptors and hetero-oligomerization with the type II receptor are needed for the antimitogenic effect⁵. Moreover, studies of a series of signalling-defective T β R-I receptors revealed that a kinase-defective T β R-I can complement an activation-defective T β R-I, suggesting that the signalling complex consists of at least two T β R-I molecules¹⁵.

In the receptor-activation model, T β R-I acts downstream of T β R-II for most, if not all, TGF- β -mediated responses, and the type I receptor thus determines the specificity of the intracellular signals¹⁶. A nine-amino-acid sequence between kinase subdomains IV and V of T β R-I, which diverges between the different type I receptors, is important for transduction of specific TGF- β signals¹⁷ (Fig. 1).

Situations have been described where, as a result of a decrease in expression of T β R-II but not T β R-I, cells lose the antiproliferative response to TGF- β , whereas the matrix accumulation induced by TGF- β is retained^{18,19}. Moreover, expression of a dominant-negative T β R-II was found to block the growth-inhibitory effect of TGF- β , but not the effect on extracellular matrix^{19,20}. These observations are compatible with a more important role for T β R-II in the anti-proliferative response than in the matrix response, but they do not necessarily contradict the sequential activation model in which activated T β R-I is required for both responses; different effects of TGF- β may occur at different threshold levels of stimulation, with the antiproliferative effect requiring a more efficient stimulus than the effect on matrix, for example.

An important step in receptor activation is phosphorylation of the tetrameric receptor complex. Phosphorylation sites in T β R-II and T β R-I have been mapped using wild-type and chimaeric receptors^{21–23} (Fig. 1). Certain of the phosphorylation sites in T β R-II are important in modulating the signalling activity of the receptor; phosphorylation of Ser 213 and Ser 409 is required for T β R-II activity, whereas phosphorylation of Ser 416 inhibits T β R-II signalling²². Notably, T β R-II and the activin type IIB receptor autophosphorylate on tyrosine residues, as well as on serine and

threonine residues, and so may function as dual-specificity kinases^{23,24}. The importance of autophosphorylation on tyrosine residues remains to be determined.

T β R-I is phosphorylated by T β R-II at several residues in the GS domain, which leads to activation of the T β R-I kinase^{2,21} (Fig. 1). It is possible that phosphorylation of T β R-II in the corresponding part of the juxtamembrane region²² is also important for its activation. In addition, T β R-I is phosphorylated at Ser 165, which is located N-terminally of the GS domain. Mutation of Ser 165 gave a type I receptor with a more powerful signalling effect in growth inhibition and matrix accumulation, but a weaker apoptotic signal²¹. Thus phosphorylation of Ser 165 may modulate TGF- β signalling.

Downstream signalling mechanisms

Recent studies in the genetically accessible *Drosophila* and *Caenorhabditis elegans* have led to a breakthrough in our understanding of how signals are transduced from serine/threonine kinase receptors to the nucleus.

In *Drosophila*, the BMP-2/4 homologue Decapentaplegic (Dpp) acts by binding to the type II receptor Punt and to the type I receptors Thick veins and Saxophone. In a genetic screen for dominant enhancers of weak *dpp* alleles, *mothers against dpp* (*Mad*) and *Medea* were discovered^{25,26}. Homozygous *Mad* mutants were found to have a phenotype similar to *dpp* mutants, with defects in midgut morphogenesis, imaginal disc development and embryonic dorsal–ventral patterning²⁶. Evidence that *Mad* is a downstream component in the Dpp pathway came from the finding that *Mad* partially rescued the eye phenotype of *dpp*^{blk27}, that *Mad* is required for the response to Dpp of the visceral mesoderm or endoderm²⁸, and that *Mad* mutations suppress dominant *thick veins* alleles²⁹. There is also biochemical evidence that *Mad* functions downstream of Dpp receptors in *Drosophila*³⁰.

In *C. elegans*, *daf-1* and *daf-4* encode serine/threonine kinase receptors. *Daf-4* mutants are dauer-constitutive and smaller than wild-type; moreover, females are defective in egg-laying and males have fused tail rays. Screening for mutants with similar phenotypes revealed three genes, *sma-2*, *sma-3* and *sma-4*, which proved to be homologous to *Mad* of *Drosophila*³¹. As *Sma-2* acts in the same cell as *Daf-4* and *daf-4* is unable to rescue *sma-2* mutations, it was concluded that *Sma* molecules are involved in downstream signalling from the *Daf-4* receptor.

SMADs are cytoplasmic mediators

At least nine genes homologous to *Mad* and *sma* have been identified in *Xenopus*, mouse and man, and shown to be components in signal transduction pathways downstream of serine/threonine kinase receptors (reviewed in ref. 1) (Fig. 2a). In an attempt to simplify the nomenclature, the designation *Smad* has been suggested for vertebrate homologues of *Sma* and *Mad*.

SMADs are molecules of relative molecular mass 42K–60K with two regions of homology at the amino and carboxy terminals, termed Mad-homology domains MH1 and MH2, respectively, which are connected with a proline-rich linker sequence (Fig. 2b). Recent work, which will be discussed below, suggests that in their inactive configurations, the MH1 and MH2 domains of SMADs make contact with each other: after activation by receptors, the molecules open up, form hetero-oligomeric complexes, and translocate to the nucleus where the transcription of target genes is affected.

Pathway-restricted SMADs. Different members of the SMAD family have different roles in signalling. Smad1, Smad2, Smad3 and possibly Smad5 interact with and become phosphorylated by specific type I serine/threonine kinase receptors and thereby act in a pathway-restricted fashion. An initial indication of a functional subspecialization among different SMADs came from the finding that *Xenopus* Smad1 (Xmad1) induces ventral mesoderm, a BMP

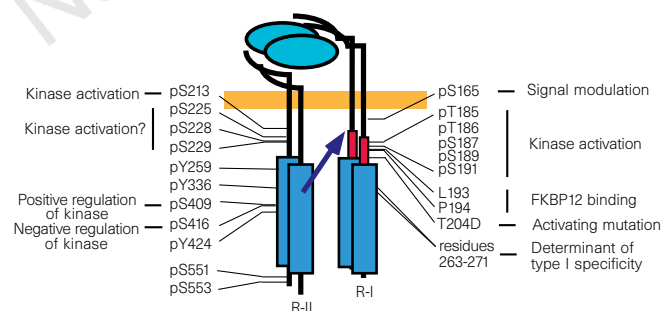


Figure 1 An activated TGF- β -receptor complex. The dimeric TGF- β molecule (light blue) binds to a heterotetrameric complex of two T β R-I and two T β R-II molecules. The GS domain (red) just upstream of the kinase domain (dark blue) of T β R-I is indicated. Known autophosphorylation sites in T β R-II and sites in T β R-I phosphorylated by T β R-II and their functional roles are indicated, as well as amino-acid residues involved in FKBP12 binding, activating mutation (T204D) and determination of T β R-I specificity in signalling (residues 263–272). Assignments of phosphorylation sites are based on data in refs 21–23. The amino-acid numbers for sites in the T β R-II in ref. 21 differ by two owing to a mistake in the amino-acid numbering in this reference.

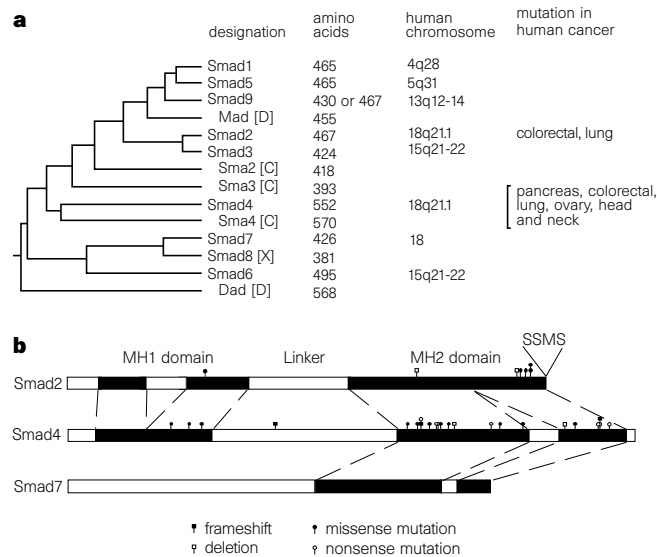


Figure 2 The SMAD family. **a**, A phylogenetic tree of human SMADs, Smad8 from *Xenopus* (X), Mad and Dad from *Drosophila* (D) and Sma from *C. elegans* (C) are shown. Alternative designations are as follows: Smad1 (Madr1, Xmad1, bsp1, Dwf-A, JV4-1), Smad2 (Madr2, Xmad2, JV18-1), Smad3 (hMad3, JV15-2), Smad4 (DPC4), Smad5 (Dwf-C, JV5-1), Smad6 (JV15-1) and Smad9 (MADH6). **b**, The sequences of representatives of pathway-restricted SMADs (Smad2), common-partner SMADs (Smad4) and inhibitory SMADs (Smad7) are shown to illustrate areas of homology between various types of SMAD molecules (black). Mutations in Smad2 and Smad4 detected in human cancers are indicated.

response^{32,33}, whereas Smad2 (Xmad2) induces dorsal mesoderm, an activin or Vg-1 response³². SMAD molecules are well conserved and act across species: human Smad1 (ref. 34), as well as *Drosophila* Mad²⁸, has ventralizing activity on *Xenopus* mesoderm, and mouse³⁵ and human³⁶ Smad2 have dorsalizing activity.

A picture is emerging in which different pathway-restricted SMADs couple to different receptors (Table 1). Smad2 and Smad3 are phosphorylated and translocated to the nucleus after stimulation by TGF- β ³⁶⁻³⁸ or activin (ref. 39 and A. Shimizu *et al.*, unpublished observation). Smad2 and Smad3 are very similar in their structures (Fig. 2a), and it is not surprising that there may be some redundancy in the functional activity between these two members of the family. Smad1 is phosphorylated and translocated into the nucleus after stimulation with BMP-2 (refs 29, 40) or BMP-4 (ref. 34). Smad5 induces ventral mesoderm in *Xenopus*⁴¹, and the recently described Smad9/MADH6 (ref. 42) is structurally similar to Smad1 and Smad5; these molecules may also be involved in BMP

signalling. There are reports that TGF- β also induces the phosphorylation of Smad1 (refs 43, 44), which may reflect redundancy or cooperativity in signalling (or a crossreactivity of antisera).

The phosphorylation of pathway-restricted SMADs by type I receptors triggers their activation. In the most C-terminal regions, pathway-restricted SMADs have a characteristic Ser-Ser-X-Ser (SSXS) motif, the two-most C-terminal serine residues of which are phosphorylated by type I receptors^{40,45-47}. Pathway-restricted SMADs bind directly to type I receptors, as demonstrated by the co-immunoprecipitation of Smad2 or Smad3 with the type I and type II receptors affinity-crosslinked with ¹²⁵I-labelled TGF- β (refs 37, 38, 45). The association between T β R-I and Smad2 or Smad3 is dependent on the kinase activity of T β R-II, but was seen only with the kinase-inactive form of T β R-I and not with wild-type T β R-I (refs 38, 45). Moreover, the phosphopeptide maps of Smad1 phosphorylated in BMP-stimulated cells were similar to those of Smad1 phosphorylated *in vitro* by purified BMP type I receptor⁴⁰. Taken together, these data suggest that the pathway-restricted SMADs are direct substrates of the type I receptor kinases, although a possible involvement of other kinases in SMAD activation has not been excluded. The data also suggest that the interaction between pathway-restricted SMADs and type I receptors is transient; presumably, SMADs are released from the receptors after phosphorylation. This idea is further supported by the observation that Smad2 molecules mutated at the three serine residues in the SSXS motif stably bind to the receptor and have dominant-negative effects⁴⁵⁻⁴⁷.

Common-mediator SMADs. The mode of action of Smad4 differs from those of other members of the SMAD family. After ligand stimulation and phosphorylation of pathway-restricted SMADs, Smad4 forms hetero-oligomers with pathway-restricted SMADs^{37,40,48,49}, which in turn translocate into the nucleus and activate transcriptional responses (Fig. 3a). In mammalian cells, Smad4 forms complexes with Smad2 and Smad3 after activation of TGF- β or activin type I receptors^{38,48,49}, whereas it forms complexes with Smad1 (refs 40, 48), and possibly with Smad5 and Smad9, after activation of BMP type I receptors. Consequently, injection of Smad4 messenger mRNA into *Xenopus* animal caps induces both ventral and dorsal mesoderm^{48,50} through the formation of complexes with Smad1, Smad5 or Smad9 and Smad2 or Smad3, respectively. Smad4, which lacks the C-terminal SSXS motif, does not bind to, nor is it phosphorylated by, TGF- β or BMP receptors^{37,38,45,48}. The phosphorylation of Smad4 has been reported to increase after activin stimulation⁴⁸, although the functional importance of this remains to be determined.

So far, only Smad4 has been identified as a common-mediator SMAD in vertebrates. A Smad4 homologue has been identified in *Drosophila* (ref. 51; and P. Das and R. W. Padgett, personal com-

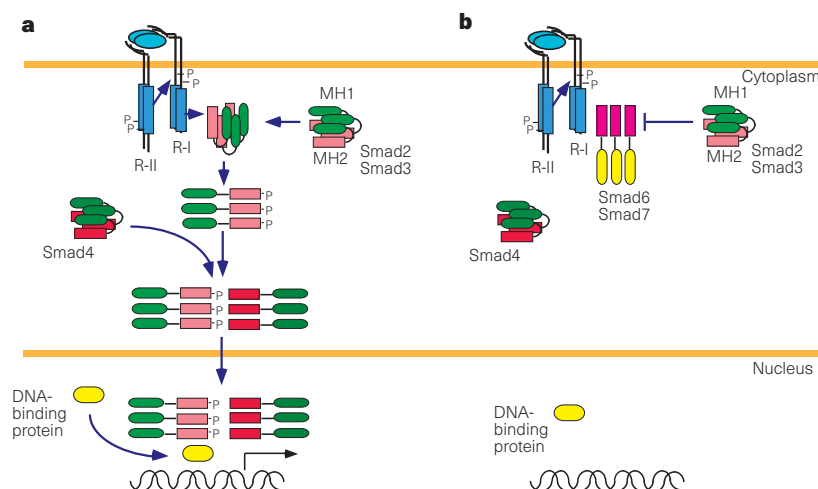


Figure 3 Agonistic and antagonistic SMAD proteins in TGF- β signalling. **a**, A hypothetical signal transduction pathway for TGF- β . TGF- β binding leads to the assembly of a heterotetrameric receptor complex in which the type II receptor phosphorylates and activates the type I receptor. Pathway-restricted SMADs (Smad2 and Smad3), which may be anchored in the cytoplasm in homotrimeric forms, are phosphorylated, which leads to heteromerization with Smad4, a common-mediator SMAD. The hetero-oligomeric complex is then translocated to the nucleus, where it binds directly or in complex with other component(s) to DNA and affects transcription of specific genes. Note that it is not known if the hetero-oligomer between Smad2, Smad3 and Smad4 is a hexamer or has another stoichiometry. **b**, Inhibitory SMADs (Smad6 and Smad7) bind to the receptors, and prevent the phosphorylation and signalling activity of pathway-restricted SMADs. Whether inhibitory SMADs occur as monomers or multimers is not known.

munication) and *C. elegans* (Sma-4). The presence of homologue of both pathway-restricted SMADs and common-mediator SMADs in lower species suggests that complex formation of these two types of SMAD molecules may be a conserved mechanism for signalling downstream of serine/threonine kinase receptors.

Functional roles of SMAD/MAD domains. Structural studies and studies of SMAD mutants have now provided an insight into the functional roles of the different domains in SMADs (summarized in Fig. 4).

The MH2 domain of SMADs may serve as an effector domain in signal transduction, as suggested by its ability to induce a transcriptional response when fused to a yeast GAL4 DNA-binding domain³⁴ and by the finding that the MH2 domain of Smad2 induces a full range of activin responses in the absence of the MH1 domain³⁵. However, in Smad4 the MH2 domain alone is not sufficient for signal transduction: a part of the linker region (termed the Smad4 activation domain⁵²) that is not conserved in other SMADs is required for signalling activity, together with the MH2 domain^{50,52}. Moreover, the MH2 domains of Smad1, Smad2 and Smad3 mediate homomeric interactions and are responsible for activation-induced interactions with Smad4 (refs 50, 53, 54).

In the resting cell, SMADs are localized in the cytoplasm. Stimulation with ligand leads to translocation to the nucleus. The observation that mouse Smad2 containing the linker region and MH2 domain localizes to the nucleus in the absence of ligand stimulation³⁵ suggests that the MH1 domain of Smad2 anchors the molecule in the cytoplasm. An alternative possibility is that truncation or ligand-induced heteromerization of SMADs leads to exposure of nuclear targeting sequence(s) in the linker region or the MH2 domain⁵⁰.

The MH1 domain of Smad4, and possibly of other SMADs, plays a role as a negative regulator by interacting with the MH2 domain, thereby preventing hetero-oligomer formation between pathway-restricted SMADs and common-mediator Smad⁵⁴. Phosphorylation of the C-terminal SSXS motif in pathway-restricted SMADs appears to remove this inhibition. Mutations in SMAD MH1 domains have been reported in certain cancers, at Arg 133 in Smad2 and at the corresponding Arg 100 in Smad4 (refs 36, 55); these mutant SMADs form homo-oligomers but cannot form Smad2–Smad4 hetero-oligomers and cannot transduce signals. This inhibition occurs by increased affinity of the mutated MH1 domains to the corresponding MH2 domains, which leads to an augmentation of the auto-inhibitory function of the MH1 domain⁵⁴.

In addition to its role as a repressor of the MH2 domain, the MH1 domain, together with part of the linker region, may also be involved in direct DNA binding as indicated by studies on *Drosophila* Mad⁵⁶. In *Drosophila*, Mad mediates the Dpp-dependent transcription of the *vestigial* (*vg*) gene through a sequence-specific DNA-binding activity. Binding of Mad to the 'quadrant' enhancer of *vg* was observed, but only when the MH2 domain was removed⁵⁶. Thus, for the binding of the MH1 domain of Mad to the *vg* quadrant enhancer, the MH2 domain appears to have a repressor function. It remains to be determined whether mammalian SMAD MH1 domains also have direct DNA-binding activity.

Three-dimensional structure of SMADs. The three-dimensional structure of the MH2 domain of Smad4 has been determined by crystallography at 2.5 Å resolution⁵⁷. It consists of a β -sandwich with antiparallel β -sheets, capped at one end by a three- α -helix bundle and at the other by three large loops and an α -helix (termed the loop/helix region). The MH2 domain of Smad4 forms a homotrimer in the crystal with the loop/helix region of one subunit interacting with the three- α -helix bundle of another. Size estimates by gel chromatography support the notion that the MH2 domain, as well as the intact Smad4 molecule, occur as trimeric structures in solution as well⁵⁷. The MH2-domain trimer is in the form of a disc, with the amino terminals of all monomers, where the MH1 domains would be attached, phasing the same side. The other side of the disc may interact with other SMAD homotrimers in the heteromeric complex. The mutations in MH2 domains detected in human cancer cells may disrupt the structure of the oligomer: some mutations disrupt the folding of the protein; others, located at the loop/helix region or the three- α -helix bundle, prevent the formation of the homotrimer; others, located at a loop (L3) in the loop/helix region which is exposed on the surface of the disc, disrupt the formation of heteromers but not of homotrimers, indicating that this region may be exposed on the surface of the disc and be critical for heteromer formation⁵⁷. In all cases, the assembly of heteromeric complex is prevented and the cell is deprived of antiproliferative TGF- β signals.

Activation of SMADs. Maximum transcriptional effect requires the cooperation between pathway-restricted SMADs and Smad4 (refs 37, 38, 48). Activation of type I receptors triggers the assembly of heteromeric complexes of the two types of SMADs, by phosphorylation of pathway-restricted SMADs in their C-terminal SSXS motifs. The mechanism may involve a phosphorylation-induced unfolding of the N- and C-terminal domains, allowing interaction with Smad4 to occur, and/or a direct interaction between the phosphorylated tail of pathway-restricted SMADs and Smad4 (ref. 46). Given the trimeric structure of Smad4 (ref. 57), such complexes may be hexamers, but their exact stoichiometry is unknown. Observations suggesting that other configurations of the active complex are possible is that full activity in a transcriptional assay can only be achieved when Smad2, Smad3 and Smad4 are all present, and that not only does Smad4 interact with Smad2 and Smad3, but Smad2 and Smad3 also interact with each other in a TGF- β -dependent manner³⁸.

Inhibitory SMADs. Smad6 and Smad7 diverge structurally from other members of the SMAD family^{58–61}; whereas they share sequence similarity with other SMADs in their C-terminal domains, their N-terminal regions (36% identical between Smad6 and Smad7) differ from those of other SMADs. Inhibitory Smads have also been detected in *Xenopus* (Smad8; J. Christian, personal communication) and *Drosophila* (Dad; ref. 62). Smad6 and Smad7 function as inhibitors of TGF- β , activin and BMP signalling. They bind to type I receptors and interfere with the phosphorylation of the pathway-restricted SMADs. Consequently, active heteromeric Smad complexes are not formed. A requirement for binding of inhibitory SMADs to type I receptors is the activation of type I

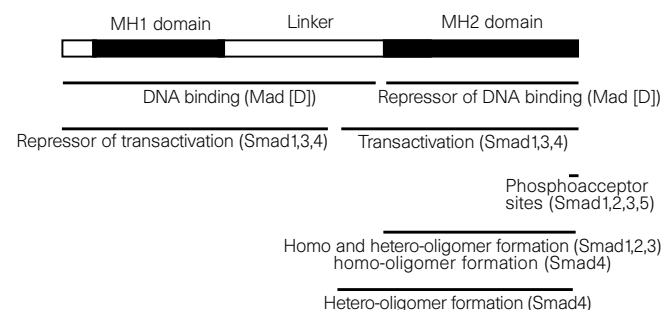


Figure 4 Different functional domains in SMADs. The conservation of MH1 and MH2 domains in the SMADs suggests that these domains may have similar functions in different members of the SMAD family, but it remains to be shown to what extent observations made on individual members (in parentheses) can be generalized.

receptor by type II receptor kinase. However, inhibitory SMADs show a more stable interaction with type I receptors than do pathway-restricted SMADs. As pathway-restricted SMADs can compete with inhibitory SMADs for binding, a plausible mechanism for inhibition is to prevent the receptor interaction and phosphorylation of pathway-restricted SMADs (Fig. 3b). In an analogous way, Dad blocks the *Drosophila* phenotype induced by activated receptor or Mad⁶², suggesting that Dad may directly interfere with the function of Mad.

Transcription of inhibitory SMAD mRNA is induced by stimulation by TGF- β as well as by other stimuli (ref. 59; and M. Kawabata *et al.*, unpublished observation), and in *Drosophila* Dad is induced by Dpp⁶². Thus, inhibitory SMADs may act as autoregulatory negative-feedback signals in the signal transduction of the TGF- β superfamily.

Transcriptional regulation by SMADs

TGF- β family members mediate their multifunctional effects by eliciting transcriptional responses on many target genes. The responsive elements in the promoters of some of these genes have been mapped and interacting transcription factors identified. In most cases, however, it is still unclear whether these genes are direct targets, and it is unknown whether SMADs are directly involved in their transcriptional regulation. Activin induces the transcription of the homeobox gene *gooseoid*⁶³ and the forkhead/winged-helix transcription factor *XFKHI* (ref. 64) without requirement of protein synthesis. Putative SMAD target genes in the BMP pathway include *Xom*⁶⁵, *Xvent-1* (ref. 66) and *Mxx-1* (ref. 67). TGF- β potentially induces transcription of plasminogen activator inhibitor-1 (ref. 68) and itself⁶⁹, with involvement of the AP-1 transcription factor, as well as the cyclin-dependent-kinase (CDK) inhibitors p15 (ref. 70) and p21 (ref. 71), with involvement of Sp1. Whether SMADs interact with AP-1 and Sp1 is not known.

Studies on the activation of *Xenopus Mix.2* and *Drosophila vg* provide the strongest evidence so far for a function of SMADs as transcriptional modulators downstream of serine/threonine kinase receptors. The homeobox gene *Mix.2* is an early-response gene induced by TGF- β superfamily members during early *Xenopus* development⁷². Smad2, Smad4 and FAST-1, a new member of the winged-helix transcription factor family, are components of an activin-responsive factor (ARF) that interacts directly in an activin-dependent manner with a 6-base-pair repeat in the activin-response element of the *Mix.2* promoter⁷². FAST-1 is the principal DNA-binding component in ARF. A C-terminal domain (amino acids 380–506) of FAST-1, termed SMAD-interacting domain (SID), is involved in binding to Smad2 and Smad4, and overexpression of SID specifically inhibits activin signalling⁷³. The C-terminal part (amino acids 453–506) of SID is essential for the association of FAST-1 with Smad2, which occurs in the absence of Smad4. Phosphorylation of Smad2 enhances the interaction between Smad2 and FAST-1. The N-terminal part (amino acids 380–453) of SID appears to be required for the interaction of Smad4 to the Smad2–FAST-1 complex, but Smad4 does not directly bind FAST-1 in the absence of Smad2. Binding of Smad4 may stabilize the Smad2–FAST-1 complex as an active DNA-binding complex. The activin-response elements in promoters of *gooseoid* from different species⁶³, and *Xenopus XFKHI/XFD-1* (ref. 64) have been mapped and show little sequence similarity with the activin-response element in the *Mix.2* gene. This suggests that different SMAD-containing transcription factor complexes can be formed which show different DNA-binding specificities.

Drosophila Mad was shown to bind to a (G + C)-rich sequence and to be essential for activation of quadrant enhancer of *vg*⁶⁶. The binding of Mad to DNA appears to be of low affinity, although it is specific, suggesting the need for heteromeric complex formation with a Smad4 homologue and/or other cofactors for high-affinity binding. This is also indicated by the finding that overexpression of

the C-terminal domain of SMADs is sufficient to mimic the effects of ligand stimulation³⁵ and suggests that the principal DNA-binding component in the transcription factor complex is not provided by a SMAD. An intrinsic low-affinity DNA-binding activity in the MH1 domain may be complemented by specific interactions with other transcription factors through the MH2 domain.

Genetic analysis of Dpp signalling in *Drosophila* has implicated Schnurri (Shn) as an essential downstream component of Dpp-dependent signalling in embryonic endoderm pattern formation^{74,75}. The *Shn* gene encodes a putative zinc-finger transcription factor with similarity to mammalian transcription factors of the major histocompatibility complex (MHC)-binding protein family. Shn activity, rather than its expression, appears to be regulated by Dpp. Thus, it is possible that Mad interacts directly with Shn to regulate transcriptional responses.

TGF- β family members may act as morphogens and induce concentration-dependent responses. These responses can be reproduced with increasing doses of SMADs^{35,76}, so it is possible that the level of nuclear SMAD provides a direct readout for the level of ligand-induced receptor activation. Promoters with different affinities for SMAD-containing transcription factor complexes may thus become activated in cells along a concentration gradient of TGF- β family members.

Receptor-interacting proteins

SMADs are clearly crucial for signal transduction of members of the TGF- β family. Using yeast two-hybrid screens, other molecules interacting with type I and type II serine/threonine kinase receptors have also been identified, which may modulate receptor signalling.

The FK506-binding immunophilin FKBP12 interacts with unstimulated T β R-I and other type I receptors. FKBP12 is not a substrate for the receptor kinase, and T β R-I mutants that are unable to interact with FKBP12 can still signal positively^{77–79}. FKBP12 binds to a Leu-Pro sequence in the GS domain of type I receptors^{77,79} (Fig. 1), and counteracts phosphorylation of the type I receptors by type II receptors; it is released from the type I receptors after ligand-induced receptor activation⁷⁸. Thus, FKBP12 protects against ligand-independent, spontaneous activation of type I receptors by type II receptors⁷⁹.

Also, the α -subunit of farnesyltransferase can interact with T β R-I^{80–82}. TGF- β stimulation does not alter the farnesyltransferase activity in mink lung cells, however, and this enzymatic activity is dispensable for the antiproliferative effect on TGF- β and its transcriptional responses in these cells⁸¹. Using a similar methodology, molecules that interact with T β R-II have also been identified—namely apolipoprotein J (ref. 83) and a WD-domain-containing protein, TRIP-1 (ref. 84), whose functional importance is not known.

Other cytoplasmic signalling pathways

In addition to the pathways already described, other parallel pathways may exist that could be important for the transduction of specific signals. Examples include TAK-1, a serine/threonine kinase of the MAP kinase kinase kinase family, which is activated by TGF- β or BMP-4 (ref. 85), and members of the Ras⁸⁶ or Rac⁸⁷ families of small GTP-binding proteins which also have been implicated in TGF- β signalling. Certain MAP kinases, such as the extracellular signal-regulated kinases (ERK)1 and 2 and stress-activated protein kinase (SAPK)/Jun-N-terminal kinase (JNK), have also been reported to be activated by TGF- β in certain cell types^{88,89}. Several of these pathways are efficiently activated in response to other signalling molecules. This may thus be yet another example of crosstalk between different signalling pathways, which appears to be common in signal transduction.

Subversion of signalling in tumorigenesis

TGF- β has a multifunctional role in tumorigenesis. At early stages,

when cells still respond to its antimitogenic effect, TGF- β may act as a tumour suppressor. However, during malignant progression, when cells acquire an insensitivity to growth inhibition by TGF- β , it may function as a tumour promoter by stimulation of angiogenesis, immunosuppression and synthesis of extracellular matrix, which provides an appropriate microenvironment for rapid tumour growth and metastasis. The biphasic action of TGF- β in tumorigenesis was demonstrated in a mouse skin model of multistage carcinogenesis using transgenic mice with keratinocyte-targeted TGF- β 1 expression⁹⁰.

The escape from the antimitogenic response of cells by TGF- β during tumour progression suggests a potential function for components in the TGF- β signal transduction pathway as tumour suppressors⁹¹. Support for a tumour-suppressor role for the type II receptor of TGF- β came from the analysis of an inherited form of colon cancer with a microsatellite instability phenotype⁹². Moreover, the frequent mutation and homozygous deletion of *Smad4* in pancreatic cancers led to its original discovery as the *DPC4* tumour-suppressor gene⁹³ (Fig. 2b). Loss of *Smad4* expression has been identified in various TGF- β -resistant cancer cells, and transfection of *Smad4* in these cells rescues responsiveness to TGF- β ^{37,52,55}. The MH2 domain of *Smad4* is often the target for point mutations and frameshift mutations that lead to premature stops. Mutations in the MH2 domain may disrupt the core structure of the protein, or perturb the ability to form stable homotrimers or hetero-oligomers with pathway-restricted SMADs, depending on which amino-acid residues are mutated⁵⁷ (see above). Somatic mutations in *Smad4* are frequently observed in pancreatic cancers⁹³, but less frequently in other types of cancers such as colon, breast and lung cancers. Functionally disruptive mutations in *Smad2*, a gene that is located close to *Smad4* on chromosome 18, have so far been noted only in colorectal and lung cancers^{36,94} (Fig. 2). *Smad1*, *Smad3*, *Smad5* and the MH2 domain of *Smad6* (*JV15-1*) do not appear to be frequently mutated in colon, breast, lung and pancreatic cancers⁹⁵. The finding of a higher frequency of somatic mutations in *Smad4* than in other *Smad* genes is consistent with a unique and non-redundant role for the common partner *Smad4* in TGF- β superfamily signalling.

TGF- β induces growth inhibition by upregulation of the CDK inhibitor p15 in certain epithelial cell lines⁹⁶. However, analysis of p15-defective human cancer cell lines reveals that the antiproliferative effect of TGF- β is also mediated by repression of the expression of *Cdc25A*, a CDK tyrosine phosphatase which activates CDK⁹⁷. *Cdc25A* is transcriptionally induced by c-Myc⁹⁸, and c-myc expression is repressed by TGF- β with similar kinetics to *Cdc25A* (refs 97, 99); an interesting possibility is therefore that c-Myc is involved in the transcriptional regulation of *Cdc25A* by TGF- β .

Conclusions

In view of their crucial importance in embryonal development, it is no surprise that signalling by TGF- β members is carefully regulated. Several intracellular control mechanisms have been discussed here. In addition, there are observations indicating that signalling is also regulated at several extracellular steps, including activation from latent precursor complexes, and interactions with specific binding proteins and accessory receptors.

Remarkable progress has been made in unravelling a signalling pathway for TGF- β and related molecules from the cell membrane all the way to the nucleus (Fig. 3). SMADs are key components in these signal transduction pathways. At present, nine vertebrate SMADs are known (Fig. 2a), but the family probably contains several other members. After phosphorylation and activation by receptor kinases, hetero-oligomeric SMAD complexes migrate into the nucleus and, either directly or in complex with other proteins, affect transcription of specific genes. Thus, the overall mechanism is reminiscent of that of signal transducers and activators of transcription (STAT) molecules, which, after phosphorylation by cytokine-

receptor-associated JAK tyrosine kinases or tyrosine kinase receptors, dimerize and move into the nucleus and induce the transcription of specific genes¹⁰⁰.

An important finding is that certain SMADs serve as inhibitors in the signal transduction of members of the TGF- β superfamily by preventing the interaction between the serine/threonine kinase receptors and pathway-restricted SMADs. As expression of the inhibitory SMADs is induced by ligand stimulation, they may have a negative-feedback role in signal transduction. Likewise, feedback switch-off signals have been found in several other signal transduction pathways and emerge as a common theme in signal transduction.

Although we are gaining our first insight into the three-dimensional structure of SMAD⁵⁷ and the involvement of SMADs in direct⁵⁶ and indirect⁷³ binding to promoter regions in specific genes, many important issues remain unresolved. Is the basis for specificity of interaction between type I receptors and different SMADs due to specific docking epitopes or to the substrate specificities of the receptor kinases? Is receptor-induced phosphorylation in the C-terminal tail sufficient for SMAD activation and initiation of all subsequent downstream signalling events, or are other phosphorylation events necessary? By which mechanism do SMADs translocate from the cytoplasm to the nucleus? Which are the partners of SMADs in the transcriptionally active complexes, and which are the target genes? Given the level of activity in the field, answers are likely to come along soon.

Note added in proof: Tyrosine kinase receptor-mediated activation of MAP kinase was recently shown to lead to phosphorylation of *Smad1* in the linker region and inhibition of its translocation to the nucleus¹⁰¹. Cross-talk between different types of signalling pathways may thus occur by differential regulation of *Smad1* activation. □

Carl-Henrik Heldin and Peter ten Dijke are at the Ludwig Institute for Cancer Research, Box 595, Biomedical Centre, S-75124 Uppsala, Sweden; Kohei Miyazono is in the Department of Biochemistry, The Cancer Institute, 1-37-1 Kamiikebukuro, Toshima-ku, Tokyo 170, Japan.

- Massagué, J., Hata, A. & Liu, F. TGF- β signalling through the Smad pathway. *Trends Cell Biol.* **7**, 187–192 (1997).
- Wrana, J. L., Attisano, L., Wieser, R., Ventura, F. & Massagué, J. Mechanism of activation of the TGF- β receptor. *Nature* **370**, 341–347 (1994).
- Henis, Y. I., Moustakas, A., Lin, H. Y. & Lodish, H. F. The types II and III transforming growth factor- β receptors form homo-oligomers. *J. Cell Biol.* **126**, 139–154 (1994).
- Chen, R.-H. & Derynck, R. Homomeric interactions between type II transforming growth factor- β receptors. *J. Biol. Chem.* **269**, 22868–22874 (1994).
- Lu, K. X. & Lodish, H. F. Signalling by chimeric erythropoietin-TGF- β receptors: Homodimerization of the cytoplasmic domain of the type I TGF- β receptor and heterodimerization with the type II receptor are both required for intracellular signal transduction. *EMBO J.* **15**, 4485–4496 (1996).
- Feng, X.-H. & Derynck, R. Ligand-independent activation of transforming growth factor (TGF) β signaling pathways by heteromeric cytoplasmic domains of TGF- β receptors. *J. Biol. Chem.* **271**, 13123–13129 (1996).
- Wieser, R., Wrana, J. L. & Massagué, J. GS domain mutations that constitutively activate TBR-I, the downstream signaling component in the TGF- β receptor complex. *EMBO J.* **14**, 2199–2208 (1995).
- Attisano, L., Wrana, J. L., Montalvo, E. & Massagué, J. Activation of signalling by the activin receptor complex. *Mol. Cell. Biol.* **16**, 1066–1073 (1996).
- Willis, S. A., Zimmerman, C. M., Li, L. & Mathews, L. S. Formation and activation by phosphorylation of activin receptor complexes. *Mol. Endocrinol.* **10**, 367–379 (1996).
- Rodriguez, C., Chen, F., Weinberg, R. A. & Lodish, H. F. Cooperative binding of transforming growth factor (TGF)- β 2 to the types I and II TGF- β receptors. *J. Biol. Chem.* **270**, 15919–15922 (1995).
- Liu, F., Ventura, F., Doody, J. & Massagué, J. Human type II receptor for bone morphogenetic proteins (BMPs): Extension of the two-kinase receptor model to the BMPs. *Mol. Cell. Biol.* **15**, 3479–3486 (1995).
- Nohno, T. et al. Identification of a human type II receptor for bone morphogenetic protein-4 that forms differential heteromeric complexes with bone morphogenetic protein type I receptors. *J. Biol. Chem.* **270**, 22522–22526 (1995).
- Rosenzweig, B. L. et al. Cloning and characterization of a human type II receptor for bone morphogenetic proteins. *Proc. Natl Acad. Sci. USA* **92**, 7632–7636 (1995).
- Yamashita, H., ten Dijke, P., Franzen, P., Miyazono, K. & Heldin, C.-H. Formation of hetero-oligomeric complexes of type I and type II receptors for transforming growth factor- β . *J. Biol. Chem.* **269**, 20172–20178 (1994).
- Weis-Garcia, F. & Massagué, J. Complementation between kinase-defective and activation-defective TGF- β receptors reveals a novel form of receptor cooperativity essential for signaling. *EMBO J.* **15**, 276–289 (1996).
- Cárcamo, J. et al. Type I receptors specify growth-inhibitory and transcriptional responses to transforming growth factor beta and activin. *Mol. Cell. Biol.* **14**, 3810–3821 (1994).
- Feng, X.-H. & Derynck, R. A kinase subdomain of transforming growth factor- β (TGF- β) type I receptor determines the TGF- β intracellular signaling specificity. *EMBO J.* **16**, 3912–3923 (1997).
- McCaffrey, T. A. et al. Decreased type II/type I TGF- β receptor ratio in cells derived from human atherosclerotic lesions. Conversion from an antiproliferative to profibrotic response to TGF- β 1. *J. Clin. Invest.* **96**, 2667–2675 (1995).

19. Sankar, S. *et al.* Modulation of transforming growth factor β receptor levels on microvascular endothelial cells during *in vitro* angiogenesis. *J. Clin. Invest.* **97**, 1436–1446 (1996).
20. Chen, R. H., Ebner, R. & Derynck, R. Inactivation of the type II receptor reveals two receptor pathways for the diverse TGF- β activities. *Science* **260**, 1335–1338 (1993).
21. Souchevsky, S., ten Dijke, P., Miyazono, K. & Heldin, C.-H. Phosphorylation of Ser165 in TGF- β type I receptor modulates TGF- β 1-induced cellular responses. *EMBO J.* **15**, 6231–6240 (1996).
22. Luo, K. X. & Lodish, H. F. Positive and negative regulation of type II TGF β receptor signal transduction by autophosphorylation on multiple serine residues. *EMBO J.* **16**, 1970–1981 (1997).
23. Lawler, S. *et al.* The type II transforming growth factor- β receptor autophosphorylates not only on serine and threonine but also on tyrosine residues. *J. Biol. Chem.* **272**, 14850–14859 (1997).
24. Nakamura, T. *et al.* Isolation and characterization of activin receptor from mouse embryonic carcinoma cells: Identification of its serine/threonine/tyrosine protein kinase activity. *J. Biol. Chem.* **267**, 18924–18928 (1992).
25. Raftery, L. A., Twombly, V., Wharton, K. & Gelbart, W. M. Genetic screens to identify elements of the decapentaplegic signaling pathway in *Drosophila*. *Genetics* **139**, 241–254 (1995).
26. Sekelsky, J. J., Newfield, S. J., Raftery, L. A., Chartoff, E. H. & Gelbart, W. M. Genetic characterization and cloning of *Mothers against dpp*, a gene required for decapentaplegic function in *Drosophila melanogaster*. *Genetics* **139**, 1347–1358 (1995).
27. Wiersdorff, V., Lecuit, T., Cohen, S. M. & Mlodzik, M. *Mad* acts downstream of Dpp receptors, revealing a differential requirement for dpp signaling in initiation and propagation of morphogenesis in the *Drosophila* eye. *Development* **122**, 2153–2162 (1996).
28. Newfield, S. J., Chartoff, E. H., Graff, J. M., Melton, D. A. & Gelbart, W. M. *Mothers against dpp* encodes a conserved cytoplasmic protein required in DPP/TGF- β responsiveness cells. *Development* **122**, 2099–2108 (1996).
29. Hoodless, P. A. *et al.* MADR1, a MAD-related protein that functions in BMP2 signaling pathways. *Cell* **85**, 489–500 (1996).
30. Newfield, S. J. *et al.* *Mothers against dpp* participates in a DPP/TGF- β responsive serine–threonine kinase signal transduction cascade. *Development* **124**, 3167–3176 (1997).
31. Savage, C. *et al.* *Caenorhabditis elegans* genes *sma-2*, *sma-3*, and *sma-4* define a conserved family of transforming growth factor β pathway components. *Proc. Natl Acad. Sci. USA* **93**, 790–794 (1996).
32. Graff, J. M., Bansal, A. & Melton, D. A. *Xenopus* Mad proteins transduce distinct subsets of signals for the TGF β superfamily. *Cell* **85**, 479–487 (1996).
33. Thomsen, G. H. *Xenopus* *mothers against decapentaplegic* is an embryonic ventralizing agent that acts downstream of the BMP-2/4 receptor. *Development* **122**, 2359–2366 (1996).
34. Liu, F. *et al.* A human Mad protein acting as a BMP-regulated transcriptional activator. *Nature* **381**, 620–623 (1996).
35. Baker, J. C. & Harland, R. M. A novel mesoderm inducer, *Mad2*, functions in the activin signal transduction pathway. *Genes Dev.* **10**, 1880–1889 (1996).
36. Eppert, K. *et al.* MADR2 maps to 18q21 and encodes a TGF β -regulated MAD-related protein that is functionally mutated in colorectal carcinoma. *Cell* **86**, 543–552 (1996).
37. Zhang, Y., Feng, X.-H., Wu, R.-Y. & Derynck, R. Receptor-associated Mad homologues synergize as effectors of the TGF- β response. *Nature* **383**, 168–172 (1996).
38. Nakao, A. *et al.* TGF- β receptor-mediated signalling through Smad2, Smad3 and Smad4. *EMBO J.* **16**, 5353–5362 (1997).
39. Chen, Y., Lebrun, J. J. & Vale, W. Regulation of transforming growth factor β - and activin-induced transcription by mammalian Mad proteins. *Proc. Natl Acad. Sci. USA* **93**, 12992–12997 (1996).
40. Kretschmar, M., Liu, F., Hata, A., Doody, J. & Massagué, J. The TGF- β family mediator Smad1 is phosphorylated directly and activated functionally by the BMP receptor kinase. *Genes Dev.* **11**, 984–995 (1997).
41. Suzuki, A., Chang, C., Yingling, J. M., Wang, X.-F. & Hemmati-Brivanlou, A. Smad5 induces ventral fates in *Xenopus* embryo. *Dev. Biol.* **184**, 402–405 (1997).
42. Watanabe, T. K. *et al.* Cloning and characterization of a novel member of the human Mad gene family. *Genomics* **42**, 446–451 (1997).
43. Lechleider, R. J., de Caestecker, M. P., Dehejia, A., Polymeropoulos, M. H. & Roberts, A. B. Serine phosphorylation, chromosomal localization, and transforming growth factor- β signal transduction by human *bsp-1*. *J. Biol. Chem.* **271**, 17617–17620 (1996).
44. Yingling, J. M. *et al.* Mammalian dwarfins are phosphorylated in response to transforming growth factor β and are implicated in control of cell growth. *Proc. Natl Acad. Sci. USA* **93**, 8940–8944 (1996).
45. Marcias-Silva, M. *et al.* MADR2 is a substrate of the TGF β receptor and its phosphorylation is required for nuclear accumulation and signaling. *Cell* **87**, 1215–1224 (1996).
46. Souchevsky, S. *et al.* Phosphorylation of Ser465 and Ser467 in the C-terminus of Smad2 mediates interaction with Smad4 and is required for TGF- β signalling. *J. Biol. Chem.* **272**, 28107–28115 (1997).
47. Abdollah, S. *et al.* TPRI phosphorylation of Smad2 on Ser465 and 467 is required for Smad2/Smad4 complex formation and signalling. *J. Biol. Chem.* **272**, 27678–27685 (1997).
48. Lagna, G., Hata, A., Hammati-Brivanlou, A. & Massagué, J. Partnership between DPC4 and SMAD proteins in TGF- β signalling pathways. *Nature* **383**, 832–836 (1996).
49. Wu, R.-Y., Zhang, Y., Feng, X.-Y. & Derynck, R. Heteromeric and homomeric interactions correlate with signaling activity and functional cooperativity of Smad3 and Smad4/DPC4. *Mol. Cell. Biol.* **17**, 2521–2528 (1997).
50. Zhang, Y., Musci, T. & Derynck, R. The tumor suppressor Smad4/DPC4 as a central mediator of Smad function. *Curr. Biol.* **7**, 270–276 (1997).
51. Padgett, R. W., Savage, C. & Das, P. Genetic and biochemical analysis of TGF β signal transduction. *Cytokine Growth Factor Rev.* **8**, 1–9 (1997).
52. de Caestecker, M. P. *et al.* Characterization of functional domains within smad4/DPC4. *J. Biol. Chem.* **272**, 13690–13696 (1997).
53. Meersseman, G. *et al.* The C-terminal domain of Mad-like signal transducers is sufficient for biological activity in the *Xenopus* embryo and transcriptional activation. *Mech. Dev.* **61**, 127–140 (1997).
54. Hata, A., Lo, R. S., Wotton, D., Lagna, G. & Massagué, J. Mutations increasing autoinhibition inactivate tumour suppressors Smad2 and Smad4. *Nature* **388**, 82–87 (1997).
55. Schutte, M. *et al.* DPC4 gene in various tumour types. *Cancer Res.* **56**, 2527–2530 (1996).
56. Kim, J., Johnson, K., Chen, H. J., Carroll, S. & Laughon, A. *Drosophila* Mad binds to DNA and directly mediates activation of *vestigial* by Decapentaplegic. *Nature* **388**, 304–308 (1997).
57. Shi, Y., Hata, A., Lo, R. S., Massagué, J. & Pavletich, N. P. A structural basis for mutational inactivation of the tumour suppressor Smad4. *Nature* **388**, 87–93 (1997).
58. Imamura, T. *et al.* Smad6 is an inhibitor in the TGF- β superfamily signalling. *Nature* **389**, 622–626 (1997).
59. Nakao, A. *et al.* Identification of Smad7, a TGF- β -inducible antagonist of TGF- β signalling. *Nature* **389**, 631–635 (1997).
60. Hayashi, H. *et al.* The MAD-related protein Smad7 associates with the TGF β receptor and functions as an antagonist of TGF β signaling. *Cell* **89**, 1165–1173 (1997).
61. Topper, J. N. *et al.* Vascular MADs: Two novel MAD-related genes selectively inducible by flow in human vascular endothelium. *Proc. Natl Acad. Sci. USA* **94**, 9314–9319 (1997).
62. Tsuneizumi, K. *et al.* *Daughters against dpp* modulates dpp organizing activity in *Drosophila* wing development. *Nature* **389**, 627–631 (1997).
63. Watabe, T. *et al.* Molecular mechanisms of Spemann's organizer formation: conserved growth factor synergy between *Xenopus* and mouse. *Genes Dev.* **9**, 3038–3050 (1995).
64. Kaufmann, E. *et al.* Antagonistic actions of activin A and BMP-2/4 control dorsal lip-specific activation of the early response gene *XFD-1'* in *Xenopus laevis* embryos. *EMBO J.* **15**, 6739–6749 (1996).
65. Ladherr, R., Mohun, T. J., Smith, J. C. & Snape, A. M. *Xom*: a *Xenopus* homeobox gene that mediates the early effects of BMP-4. *Development* **122**, 2385–2394 (1996).
66. Gawanaka, V., Delius, H., Hirschfeld, K., Blumenstock, C. & Niehrs, C. Antagonizing the Spemann organizer: role of the homeobox gene *Xvent-1*. *EMBO J.* **14**, 6268–6279 (1995).
67. Suzuki, A., Ueno, N. & Hemmati-Brivanlou, A. *Xenopus* *msx1* mediates epidermal induction and neural inhibition by BMP4. *Development* **124**, 3037–3044 (1997).
68. Keeton, M. R., Curriden, S. A., van Zonneveld, A. J. & Loskutoff, D. J. Identification of regulatory sequences in the type I plasminogen activator inhibitor gene responsive to transforming growth factor β . *J. Biol. Chem.* **266**, 23048–23052 (1991).
69. Kim, S. J. *et al.* Autoinduction of transforming growth factor β 1 is mediated by the AP-1 complex. *Mol. Cell. Biol.* **10**, 1492–1497 (1990).
70. Datto, M. B., Yu, Y. & Wang, X.-F. Functional analysis of the transforming growth factor β responsive elements in the WAF1/Cip1/p21 promoter. *J. Biol. Chem.* **270**, 28623–28628 (1995).
71. Li, J.-M., Nichols, M. A., Chandrasekharan, S., Xiong, Y. & Wang, X.-F. Transforming growth factor β activates the promoter of cyclin-dependent kinase inhibitor p15^{INK4B} through an Sp1 consensus site. *J. Biol. Chem.* **270**, 26750–26753 (1995).
72. Chen, X., Rubock, M. J. & Whitman, M. A transcriptional partner for MAD proteins in TGF- β signalling. *Nature* **383**, 691–696 (1996).
73. Chen, X. *et al.* Smad4 and FAST-1 in the assembly of activin-responsive factor. *Nature* **389**, 85–89 (1997).
74. Arora, K. *et al.* The *Drosophila* *schnurri* gene acts in the Dpp/TGF- β signaling pathway and encodes a transcription factor homologous to the human BMF family. *Cell* **81**, 781–790 (1995).
75. Grieder, N. C., Nellen, D., Burke, R., Basler, K. & Affolter, M. *schnurri* is required for *Drosophila* Dpp signaling and encodes a zinc finger protein similar to the mammalian transcription factor PRDII-BF1. *Cell* **81**, 791–800 (1995).
76. Wilson, P. A., Lagna, G., Suzuki, A. & Hemmati-Brivanlou, A. Concentration-dependent patterning of the *Xenopus* ectoderm by BMP4 and its signal transducer Smad1. *Development* **124**, 3177–3184 (1997).
77. Chang, M.-J., Kinnunen, P., Hawker, J., Brand, T. & Schneider, M. D. FKBP-12 recognition is dispensable for signal generation by type I transforming growth factor- β receptors. *J. Biol. Chem.* **271**, 22941–22944 (1996).
78. Wang, T. *et al.* The immunophilin FKBP12 functions as a common inhibitor of the TGF β family type I receptors. *Cell* **86**, 435–444 (1996).
79. Chen, Y. G., Liu, F. & Massagué, J. Mechanism of TGF β receptor inhibition by FKBP12. *EMBO J.* **16**, 3866–3876 (1997).
80. Kawabata, M., Imamura, T., Miyazono, K., Engel, M. E. & Moses, H. L. Interaction of the transforming growth factor- β type I receptor with farnesyl-protein transferase- α . *J. Biol. Chem.* **270**, 29628–29631 (1995).
81. Ventura, F., Liu, F., Doody, J. & Massagué, J. Interaction of transforming growth factor- β receptor I with farnesyl-protein transferase- α in yeast and mammalian cells. *J. Biol. Chem.* **271**, 13931–13934 (1996).
82. Wang, T. *et al.* The p21^{RAS} farnesyltransferase α subunit in TGF- β and activin signaling. *Science* **271**, 1120–1122 (1996).
83. Reddy, K. B., Karode, M. C., Harmony, J. A. K. & Howe, P. H. Interaction of transforming growth factor β receptors with apolipoprotein J/clusterin. *Biochemistry* **35**, 309–314 (1996).
84. Chen, R.-H., Miettinen, P. J., Maruoka, E. M., Choy, L. & Derynck, R. AWD-domain protein that is associated with and phosphorylated by the type II TGF- β receptor. *Nature* **377**, 548–552 (1995).
85. Yamaguchi, K. *et al.* Identification of a member of the MAPKKK family as a potential mediator of TGF- β signal transduction. *Science* **270**, 2008–2011 (1995).
86. Hartsough, M. T. *et al.* Altered transforming growth factor β signaling in epithelial cells when Ras activation is blocked. *J. Biol. Chem.* **271**, 22368–22375 (1996).
87. Mucci, L., Skorecki, K. L. & Goldberg, H. J. Extracellular signal-regulated kinase and the small GTP-binding protein, Rac, contribute to the effects of transforming growth factor- β 1 on gene expression. *J. Biol. Chem.* **271**, 16567–16572 (1996).
88. Atfi, A., Djelloul, S., Chastre, E., Davis, R. R. & Gespach, C. Evidence for a role of Rho-like GTPases and stress-activated protein kinase/c-Jun N-terminal kinase (SAPK/JNK) in transforming growth factor beta-mediated signaling. *J. Biol. Chem.* **272**, 1429–1432 (1997).
89. Frey, R. S. & Mulder, K. M. Involvement of extracellular signal-regulated kinase 2 and stress-activated protein kinase Jun N-terminal kinase activation by transforming growth factor β in the negative growth control of breast cancer cells. *Cancer Res.* **57**, 628–633 (1997).
90. Cui, W. *et al.* TGF β 1 inhibits the formation of benign skin tumors, but enhances progression to invasive spindle carcinomas in transgenic mice. *Cell* **86**, 531–542 (1996).
91. Markowitz, S. D. & Roberts, A. B. Tumor suppressor activity of the TGF- β pathway in human cancers. *Cytokine & Growth Factor Rev.* **7**, 93–102 (1996).
92. Markowitz, S. *et al.* Inactivation of the type II TGF- β receptor in colon cancer cells with microsatellite instability. *Science* **268**, 1336–1338 (1995).
93. Hahn, S. A. *et al.* DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1. *Science* **271**, 350–353 (1996).
94. Riggins, G. J. *et al.* Mad-related genes in the human. *Nature Genet.* **13**, 347–349 (1996).
95. Riggins, R. G., Kinzler, K. W., Vogelstein, B. & Thiagalingam, S. Frequency of *Smad* gene mutations in human cancers. *Cancer Res.* **57**, 2578–2580 (1997).
96. Hannon, G. J. & Beach, D. p15^{INK4B} is a potential effector of TGF- β -induced cell-cycle arrest. *Nature* **371**, 257–261 (1994).
97. Iavarone, A. & Massagué, J. Repression of the CDK activator Cdc25A and cell-cycle arrest by cytokine TGF- β in cells lacking the CDK inhibitor p15. *Nature* **387**, 417–422 (1997).
98. Galaktionov, K., Chen, X. & Beach, D. Cdc25 cell-cycle phosphatase as a target of c-Myc. *Nature* **382**, 511–517 (1996).
99. Pietenpol, J. A. *et al.* TGF- β 1 inhibition of c-myc transcription and growth in keratinocytes is abrogated by viral transforming proteins with PRB binding domains. *Cell* **61**, 777–785 (1990).
100. Ihle, J. N. STATs: Signal transducers and activators of transcription. *Cell* **84**, 331–334 (1996).
101. Kretschmar, M., Doody, J. & Massagué, J. Opposing BMP and EGF signalling pathways converge on the TGF- β family mediator Smad1. *Nature* **389**, 618–622 (1997).

Acknowledgements. We thank I. Schiller for help in the preparation of this manuscript. We apologize to all those authors whose work we have been unable to cite because of space constraints.

Size and morphology of the Chicxulub impact crater

Jo Morgan*, Mike Warner* and the Chicxulub Working Group, John Brittan*, Richard Buffler†, Antonio Camacho‡, Gail Christeson†, Paul Denton§, Alan Hildebrand||, Richard Hobbs¶, Hamish Macintyre*, Graeme Mackenzie§, Peter Maguire§, Luis Marin#, Yosio Nakamura†, Mark Pilkington||, Virgil Sharpton☆, Dave Snyder¶, Gerardo Suarez# & Alberto Trejo#

* Department of Geology, Imperial College, London, SW7 2BP, UK

† University of Texas Institute for Geophysics, Austin, Texas 78759-8500, USA

‡ Petroleos Mexicanos, Villahermosa, CP86030, Mexico

§ University of Leicester, Leicester, LE1 7RH, UK

|| Geological Survey of Canada, Ottawa, K1A 0Y3, Canada

¶ BIRPS, Bullard Laboratories, University of Cambridge, CB3 0EZ, UK

Universidad Nacional Autonoma de Mexico, Mexico DF 04510, Mexico

☆ Lunar and Planetary Institute, Houston, Texas 77058, USA

The Chicxulub impact in Mexico has been linked to the mass extinction of species at the end of the Cretaceous period. From seismic data collected across the offshore portion of the impact crater, the diameter of the transient cavity is determined to be about 100 km. This parameter is critical for constraining impact-related effects on the Cretaceous environment, with previous estimates of the cavity diameter spanning an order of magnitude in impact energy. The offshore seismic data indicate that the Chicxulub crater has a multi-ring basin morphology, similar to large impact structures observed on other planets, such as Venus.

On the basis of an anomalous clay layer observed at the Cretaceous/Tertiary (K/T) boundary, Alvarez *et al.*¹ proposed that there had been a large impact at the end of the Cretaceous period (65 Myr ago). The Chicxulub crater in Mexico²⁻⁷, which is the largest known Phanerozoic impact structure, is now widely accepted as the site of this event. The crater is located partly offshore; onshore its main surface expression (Fig. 1) is a roughly 165 km diameter semicircular ring of cenotes⁸ (sink-holes in the carbonate platform). The crater is covered by post-impact Tertiary sediments that thicken from a few hundred metres outside the cenote ring to more than a kilometre in the interior of the basin. The crater has a clear gravitational (Fig. 1) and magnetic signature⁹⁻¹².

The formation of simple bowl-shaped impact craters of less than a few kilometres in diameter is well understood from field observations, simulations of impacts in laboratory experiments, and theoretical calculations¹³. As craters increase in size, they undergo gravity-driven modification where the floor of the initial transient cavity rebounds upwards, and the crater margins collapse inwards, to form broad, shallow, complex craters. The smallest complex craters have central peaks. As crater size increases further, this central peak is replaced by a peak ring, typically an irregular ring of hills and massifs, that lacks prominent asymmetric bounding scarps. The largest craters, at least on the Moon and Venus, appear as multi-ring basins. These may contain a peak ring, and, by definition, contain two or more prominent rings showing inward-facing asymmetric scarps. The mechanism of ring-formation in multi-ring basins remains obscure, not least because of the lack of clear terrestrial analogues and the difficulty of making high-resolution subsurface observations on other planets.

On Earth the transition from central-peak to peak-ring morphology occurs at a diameter of about 25 km (refs 13, 14). On Venus, which is comparable to the Earth in terms of its gravity, the transition occurs at around the same diameter, and the transition from peak-ring to multi-ring morphology occurs between about 110 and 150 km diameter¹⁴. We might therefore expect that craters on Earth greater than about 150 km in diameter would also have a

multi-ring basin morphology¹², although it remains to be proven whether such an expectation is justified, as the transition may be rheologically rather than gravitationally controlled. There are three craters of >150 km diameter known on Earth: Vredefort, Sudbury and Chicxulub. The morphology of the first two remains equivocal;

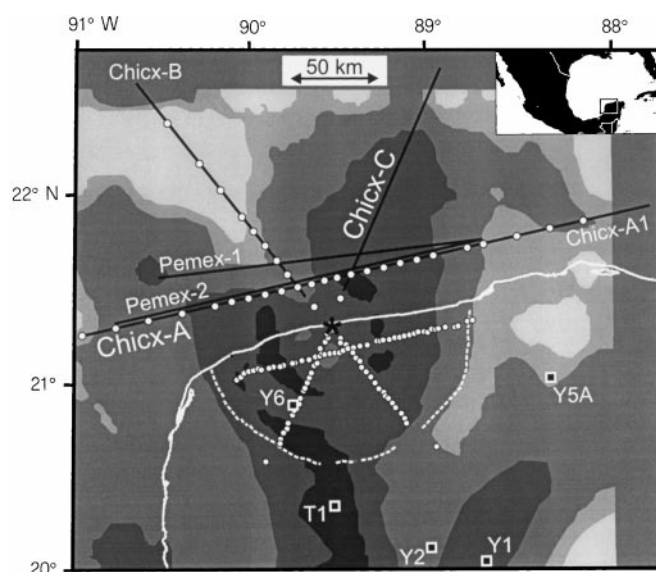


Figure 1 The Chicxulub seismic experiment. Solid lines show offshore reflection lines, white dots show wide-angle receivers. Shading shows Bouguer gravity anomaly; the crater is marked by a ~30 mGal circular gravity low. The dashed white line marks the position of the cenote ring. Squares show well locations; Y6 is ~1.6 km deep and T1, Y1, Y2 and Y5a are 3–4 km deep. We calculated all radii using a nominal centre at 89.54° W, 21.3° N, located by an asterisk. There is ambiguity in defining an exact centre as the inner and outer gravity structures, the cenote ring and the magnetic data all have slightly different centres^{8,11,12}.

they are both around 2 Gyr old, Vredefort is deeply eroded and Sudbury is tectonically deformed. Chicxulub therefore represents our best opportunity to understand the dynamics and structure of large terrestrial craters.

The precise size and morphology of the Chicxulub crater has been in dispute; it has been interpreted as a ~180 km peak-ring crater^{2,9,11}, a ~250 km peak-ring crater^{5,8}, and a ~300 km multi-ring basin^{4,12,15}. This size difference could represent an order of magnitude difference in impact energy with quite different consequences for potential environmental perturbation. Our new seismic data image the offshore portion of the structure from the crater floor to the base of the crust. These data show why previously there has been ambiguity in determining crater dimensions. The radial extent of the inner edge of the collapsed transient cavity is confirmed to have been close to the minimum estimates previously proposed⁹. However, the outer structure seems to extend further than might be expected on the basis of this small inner cavity, because the crater, at least offshore, appears to have a multi-ring basin morphology. Elements of both end-member models^{4,11} turn out to have been substantially correct.

The seismic reflection data

The Chicxulub impact occurred on a carbonate platform, into shallow water, in an area otherwise tectonically and magmatically inactive. The crater floor was subsequently buried beneath about 1 km of Tertiary carbonates, and remains relatively pristine and

undeformed. Its shallow burial, and location partly offshore, make Chicxulub an ideal target for seismic investigation. In October 1996 the British Institutions Reflection Profiling Syndicate (BIRPS) acquired ~650 km of marine seismic reflection profile (Fig. 1) across the crater, recorded to 18 s two-way travel-time (TWTT). The reflection data were recorded on a 6-kilometre array of towed hydrophones, and the airgun shots fired to generate these reflection data were also recorded at large offsets on land and on the sea bed, providing high-resolution velocity control throughout the crust. These seismic data, together with the existing gravity, magnetic and well data, provide the clearest image yet of the crater in three dimensions.

Here we describe the main features observed on the deep reflection profiles (Chicx-A, A1, B and C), and on two shallow industry profiles (Pemex 1 and 2) acquired by Petroleos Mexicanos in 1992 (ref. 6). Pemex 2, which is close to Chicx-A, was reprocessed as part of this project.

Target stratigraphy. The seismic reflection data image the pre-impact target stratigraphy, labelled A in Fig. 2, as a series of bright, continuous, subhorizontal layered reflectors between 0.5 and 2.5 s TWTT around the periphery of the crater. These reflectors can be traced as they deepen towards the crater centre and disappear at a diameter of about 85 km and a depth of about 3.5 s (7–9 km), B in Fig. 2. We assume that material originally inside this diameter has been excavated, and that the disappearance of this stratigraphy marks the inner edge of the collapsed transient cavity.

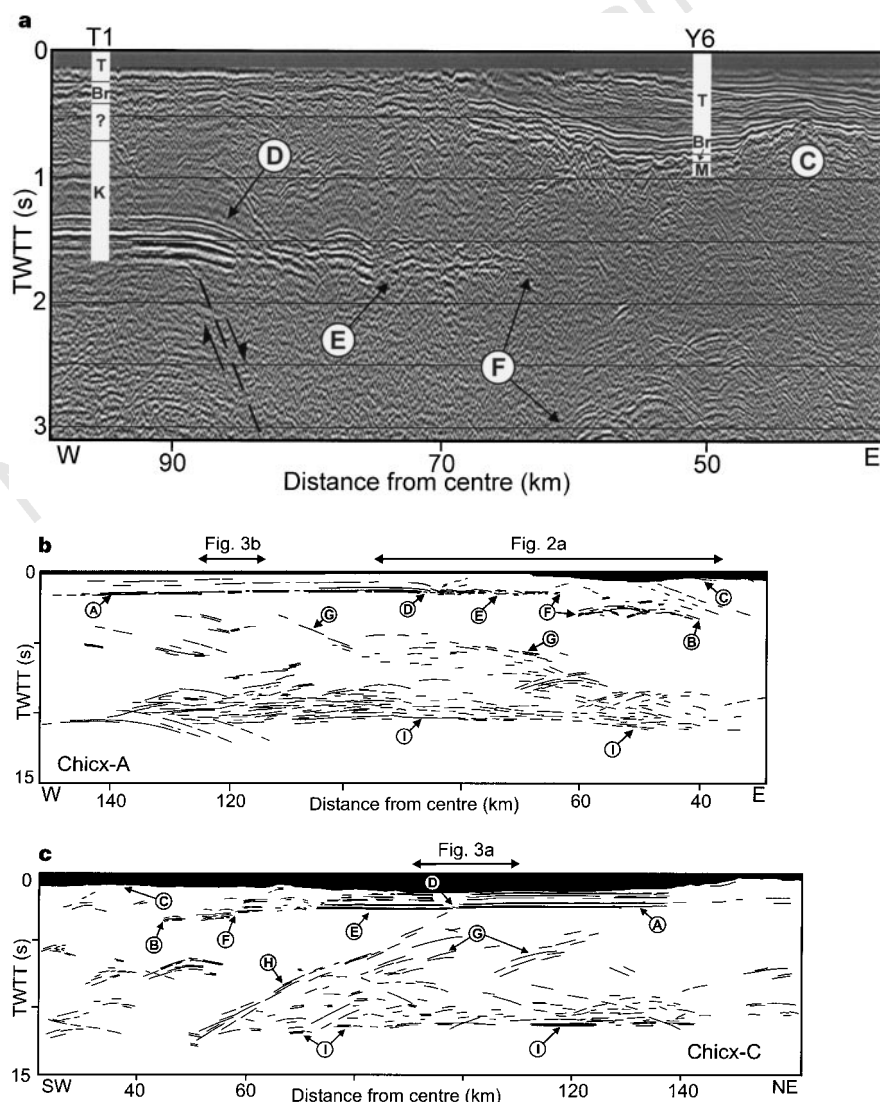


Figure 2 Seismic reflection data along part of Chicx-A and Chicx-C; the data are unmigrated. **a**, Close-up Chicx-A. **b**, Line drawing of Chicx-A. **c**, Line drawing of Chicx-C. Highly reflective target stratigraphy, labelled (A), lies at ~1.6 s on Chicx-A, and at ~2.5 s on Chicx-C. This stratigraphy deepens by 200 ms across a monocline (D), remains subhorizontal for ~20 km (E), and then slumps (F) to between 3 and 4 s to form a series of slumped blocks before disappearing at the inner edge of the collapsed transient cavity (B). The peak ring is labelled (C), the suite of dipping reflectors (G and H), and the Moho reflector at the base of the crust (I). In **a** we have projected the stratigraphy in wells Y6 and T1 on to the seismic section at their respective radial distances; T represents Tertiary rocks, Br impact breccia, K the pre-impact Cretaceous stratigraphy, and M melt breccia⁷¹².

In Fig. 2a we show the stratigraphy encountered in onshore well T1 projected onto the seismic line at its equivalent radius. No wells have yet been drilled offshore. In T1 the Tertiary sequence is ~500 m thick, and is underlain by an impact breccia interpreted as being between ~400 m (refs 7, 11) and ~1,300 m thick¹². The range in breccia thickness reflects the difficulty, when using well logs and a small number of core samples, in distinguishing between disrupted autochthonous target rock and allocthenic impact breccia. On the seismic data we observe a chaotic sequence from 0.4 to 0.9 s (0.5–1.7 km). It is likely that a significant portion of this sequence is the impact breccia observed in T1. The chaotic sequence is underlain by about 0.8 s (~2 km) of weakly disturbed layered reflections originating from the Mesozoic section that is observed in nearby onshore wells T1, Y1, Y2 and Y5a (ref. 6). The bright reflective layer at ~1.6 s (~2.5 s on Chicx-C) is most probably generated by the lower Cretaceous anhydrites interbedded with carbonates that are observed in the onshore wells.

Tertiary basin and peak ring. In the central region of the crater, we observe a ~1 s deep Tertiary basin which has a radius of 68–82 km except to the northeast where the deep basin extends to radial distances of about 140 km (Fig. 2c). The region of anomalously thick Tertiary sedimentation in the northeast coincides with a region of anomalously low gravity (Fig. 1). The stratigraphy encountered in Y6 (refs 7, 12) is projected on to the seismic line in Fig. 2a. Tertiary sediments in Y6 coincide with a layered reflective sequence in the seismic data, whereas the breccias correspond to an unreflective layer. The Tertiary basin contains an outer annular trough, and a peak ring (C) with an average radius of 40 km measured to the highest peak. The peak ring is irregular, rugged, and stands a few hundred metres above the basin floor. The ring is narrow and prominent (400–600 m above crater floor) in the west and northwest (Fig. 2a and b), and broader and less prominent (200–300 m above crater floor) to the east and northeast (Fig. 2c). On several profiles, the peak ring lies directly over the inner edge of the collapsed transient cavity (Fig. 2b).

Offsets in Mesozoic strata. The pre-impact stratigraphy is most clearly imaged on lines Chicx-A and C (Fig. 2). Tracking these reflectors towards the crater centre we observe an offset (D) in the Mesozoic sequence at a radial distance of 101 km on Chicx-C and 87 km on Chicx-A. In both cases the vertical offset is 400–500 m and it is observed in the Cretaceous stratigraphy but not in the deepest

Tertiary reflectors. On Chicx-A the offset occurs entirely across a monocline (Fig. 2a); on Chicx-C the offset occurs partly across a monocline and partly across a fault-bounded asymmetric graben (Fig. 3a). Inside these offsets, the stratigraphy remains subhorizontal, although internally disturbed, for at least 20 km (E). On Chicx-A the target stratigraphy is down-thrown by 4–5 km on a single fault (F) at a radial distance of 61 km to form a series of slumped blocks, whereas on Chicx-C we observe a sequence of faults with a combined throw of ~3 km. Pemex 1 shows a similar profile to Chicx-A at comparable radial distances⁶, except that the inner slumping occurs over several faults as it does along Chicx-C.

To the northwest and east, on Chicx-B and A1, the seismic signal is strong within the region of the deep Tertiary basin, but outside this the pre-impact stratigraphy is discontinuous and not strongly reflective. We cannot use these lines to unequivocally identify offsets in the Mesozoic sequence outside the deep Tertiary basin. On Chicx-B the region of missing Cretaceous reflectivity corresponds roughly to the position and extent of the anomalous gravity high observed in the northwest sector of the crater (Fig. 1). On Chicx-A1 problems with near-surface statics appear to be responsible for the poor shallow seismic image. On this line, we do however image clearly the main normal fault (F) that bounds the slumped blocks at a radial distance of 62 km.

On Chicx-A, the shallowest pre-impact stratigraphy becomes increasingly disturbed towards the crater centre. This disturbance is a likely result of structural collapse, secondary cratering and near-surface fracturing. Such fracturing can be caused when tensile stresses produced by the rarefaction wave locally exceed overburden pressure and the tensile strength of the rocks. On Chicx-C, the uppermost Cretaceous stratigraphy seems to be absent, removed presumably during or shortly after impact by collapse into deep water to the northeast. On Chicx-A and Chicx-A1 the upper part of the Mesozoic sequence is disrupted at a radial distance of about 120 km (Fig. 3b). On both lines this disruption appears to be more easily explained by outwardly directed thrusting than by inward collapse, but its precise significance remains unclear, and it is possible that both processes were involved.

Crustal reflectivity. The deep crust on all the profiles is highly reflective. Over much of the lines we observe a bright, subhorizontal reflector (I) at the base of this reflectivity at around 11 s which we interpret as the crust–mantle boundary. We see reflections from the

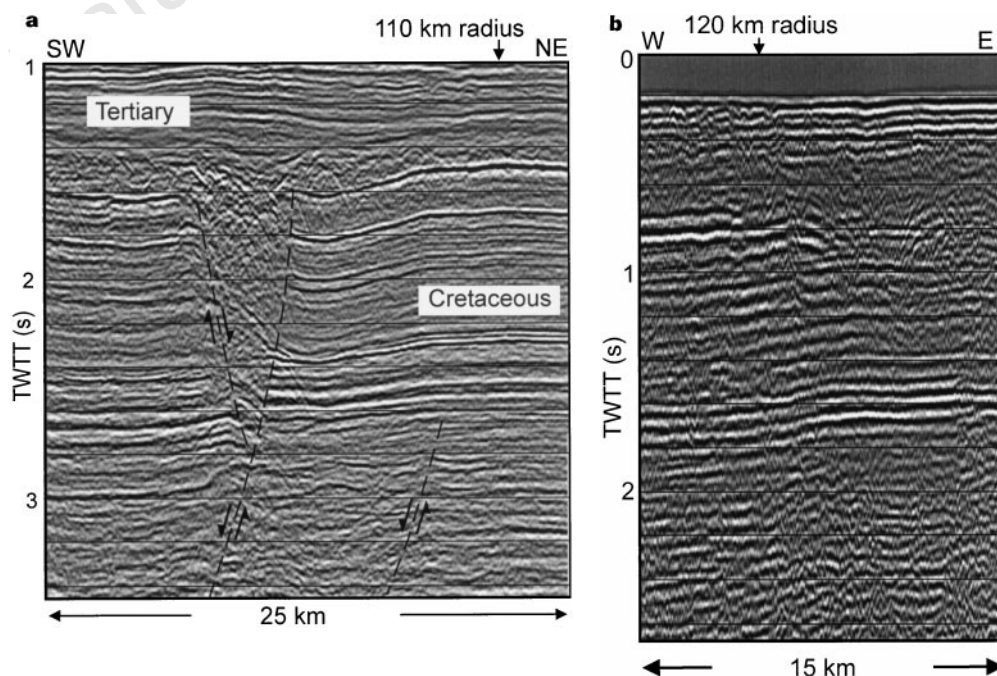


Figure 3 Seismic sections showing deformation of the target stratigraphy. **a**, Close-up of the outer ring on Chicx-C formed by a monocline and fault-bounded graben. **b**, Close-up of the disruption in the target stratigraphy at ~120 km radius on Chicx-A.

Moho discontinuity on our wide-angle data at ~ 35 km depth that correspond to these normal-incidence reflections. Bright lower-crustal reflectivity is commonly observed on deep reflection lines¹⁶, and is unlikely to be generically related to the impact. It does appear, however, that the geometry of the lower crustal reflectivity has been affected by the impact.

On all the deep reflection profiles, we observe distinctive bands of dipping, linear reflections, G and H, in the crystalline crust outside the collapsed transient cavity. These reflections dip towards the crater centre at between 30 and 40° , and extend to radial distances of 135 km. On Chicx-C the innermost reflector H in Fig. 2c, can be traced directly to the normal fault at 101 km (Fig. 3a), and, at depth, the reflections appear to offset the Moho (I) by ~ 1 s. Observations of linear reflectors that offset unequivocally both near-surface sediments and the Moho are extremely rare on continental deep seismic reflection profiles. The dipping reflections show no evidence that they dip less steeply at depth, nor do they intersect or detach into the collapsed transient cavity.

Transient and excavation cavities

The transient cavity is produced during the compressive stage of impact immediately before the gravitational collapse that leads to the formation of the final crater. The excavation cavity is formed by the boundary between material that is ejected from the crater and material that is displaced to form the transient cavity. In Fig. 4 we

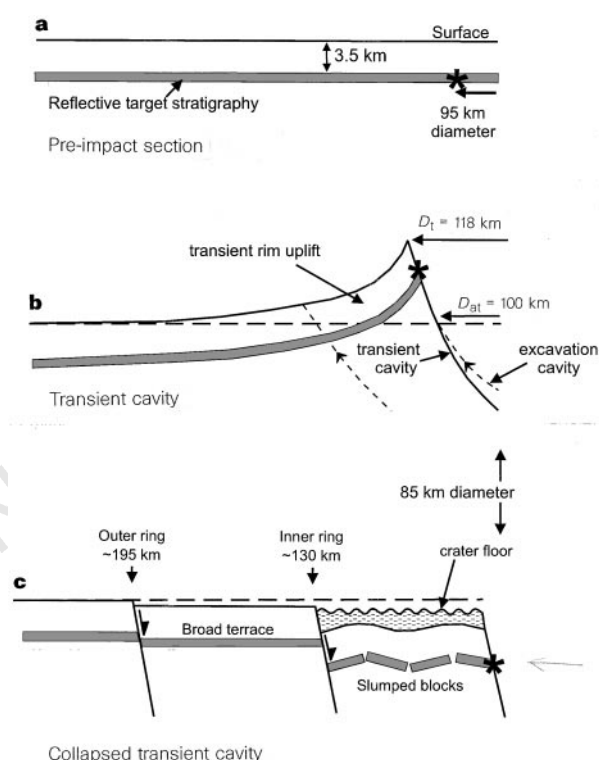


Figure 4 Reconstruction of the rim of the transient cavity. Cartoon of **a**, the pre-impact section; **b**, the transient cavity; and **c**, the collapsed transient cavity. The pre-impact surface in **a** is shown as a horizontal dashed line in **b** and **c**. The crater floor, which is ~ 1 km below the pre-impact surface, is indicated by a wavy line. The short-dashed lines in **b** represent flow lines calculated using the Z-model; the inner flow line delineates the excavation cavity. The asterisk marks reflective target rock, shaded dark grey, that lay initially just outside the excavation zone, and, after the impact, lies near the inner edge of the collapsed transient cavity. The final diameters are average values obtained from Chicx-A and Chicx-C showing rings at ~ 130 km and ~ 195 km diameter at surface. Values from Chicx-B and Chicx-A1 have not been used in these averages because the outer ring is equivocal on these lines. The dashed region contains impact deposits that form the floor of the crater.

reconstruct the pre-impact section and the transient cavity. The asterisk marks material that lay just outside the excavated cavity, was compressed in the transient rim uplift, and, after collapse, lay on the inner edge of the collapsed transient cavity. The diameter of the transient cavity can be usefully defined in one of two ways—either by the diameter measured at the pre-impact surface, or by the rim-to-rim diameter of the transient rim uplift. The former is often referred to as the *apparent* diameter (D_{at} in Fig. 4), and the latter as the *true* diameter (D_t in Fig. 4).

On our seismic data we observe the inner edge of the collapsed transient cavity (B). To determine from this the dimensions of the uncollapsed transient cavity, we must restore this stratigraphy to its original pre-impact position and estimate the flow regime that carried this pre-impact stratigraphy outwards and upwards to form the transient rim uplift. For the purpose of estimating the dimensions of the transient and excavation cavities we have assumed a Maxwell Z-model. This is an analytical model of the excavation flow, and provides a simplified kinematic description of the cratering flow field¹⁷. We have used a Z value of 2.7, a value successfully applied at Ries¹⁸ and other craters¹³, and an effective depth of burst 8 km below the surface. Our calculations are not very sensitive to realistic variations in these parameters. We also assume a paraboloid form for the transient cavity. In Fig. 4b we show two flow lines calculated from this Z-model: the inner one defines the excavation cavity.

The first stage in the reconstruction is to restore the slumped blocks to their pre-impact position. We assume that the collapse of the transient cavity (from Fig. 4b to c) must occur towards and into the cavity in such a way as to reduce its volume. Restoring the slumped blocks to their original depth places the inner edge of the reflective stratigraphy at the approximate location of the excavated cavity, as measured ~ 3.5 km below the pre-impact surface. The second stage is to project the position of the excavated cavity to the pre-impact surface (a few hundred metres below the present-day surface), and to reconstruct the transient rim uplift using the flow model. We obtain a value for the rim-to-rim diameter of the transient rim uplift, D_t , of 118 km, and a diameter for the excavation cavity of 100 km. This diameter must be close to that of the transient cavity measured at surface, D_{at} . There are various sources of error in this reconstruction; these allow extreme values for D_{at} to lie within the range 90 – 105 km, and for D_t to lie within the range 105 – 125 km. Note that our rim-to-rim diameter does not include any contribution from overlying ejecta.

Morphology

We recognize that any topographic highs at surface would have been subject to rapid erosion^{9,12}, and, because the Yucatan was immersed in a shallow sea, we would expect considerable re-distribution of the ejecta by high-energy wave action following the impact. We have therefore used major offsets in the deeper target stratigraphy to infer post-impact surface topography, accepting that this topography might have existed for only a short time. We observe a separation (E) between an inner zone of intense slumping and an outer isolated monocline (D) on Chicx-A, Chicx-C and Pemex-1. This separation is much larger than the widths of individual slump terraces in central-peak and peak-ring craters¹³, and is comparable to the spacing of rings in multi-ring basins. This suggests that, immediately after the impact, there were two distinct inward-facing asymmetric scarps along these profiles. We conclude therefore that Chicxulub had a multi-ring basin morphology, at least in the limited region where the seismic data and pre-existing stratigraphy are such that we would be able to identify such a structure. We have as yet no seismic reflection data around the entire southern half of the structure, and our seismic data elsewhere offshore are equivocal. It remains to be shown therefore whether these asymmetric scarps are true rings that extend around the entire crater.

Assuming that the seismic data are truly representative of the entire structure, then Chicxulub would have had at least three rings

in total, a peak ring with an average diameter of 80 km, an inner ring with a diameter of about 130 km at surface, and an outer ring with a diameter of about 195 km at surface (Fig. 4c). The peak ring rose a few hundred metres above a relatively flat basin floor. Our best but poorly constrained estimates suggest that the inner ring would have had about half a kilometre of throw visible at surface above the crater floor, and the outer ring slightly less. After the impact, this structure would have appeared similar to the smaller 140-km-diameter multi-ring basin Klenova on Venus, which appears to show all the same structural elements. Klenova has a similar degree of lateral and radial variability¹⁴ to that which we observe at Chicxulub. The apparent continuation of deep deformation to large radial distances, and disruptions in the stratigraphy at ~120 km radius on Chicx-A and Chicx-A1, hint perhaps at an additional ring outside 195 km diameter.

Implications

In principle, it is possible to calculate the energy of the impact from the size of the transient cavity. In practice, the required scaling relationships have large errors, particularly for large impacts. We use the Schmidt–Holsapple Pi-group scaling law¹⁹ to calculate the size of the bolide. The impact energy is about 5×10^{23} J. For an asteroid impact the required diameter is about 12 km; if the object was a comet then the required diameter lies in the range 10–14 km depending on impact velocity.

Our estimates of the size of the transient and excavated cavity place constraints on the quantities of material ejected by the impact. Using the Z-model with an effective depth of burst at 8 km, we find that the total volume of ejecta is ~80,000 km³. This is probably too high because the Z-model overestimates the depth of excavation when the depth of burst is placed below the surface¹⁷. The volume is likely to be closer to 50,000 km³, calculated assuming an effective depth of burst at surface. The maximum uplift of the transient crater rim was about 8 km. Scaling laws¹³ place the depth of the transient cavity at between 35 and 40 km, and the maximum depth of excavation at about 12 km.

Previous calculations^{20,21} of the total sulphur released into the atmosphere by the impact range from 6.5×10^{12} to 4.2×10^{15} kg. These values, which are important for modelling the environmental consequences of the impact, are poorly constrained because of uncertainties in crater size, the total volume of anhydrite in the target sequence, and the precise pressure at which sulphur is released. Using the calculations presented by Ivanov *et al.*²² together with an apparent transient diameter of 100 km, we find that the total sulphur release was between 6×10^{13} and 1.5×10^{14} kg. These volumes are not sufficient to produce the dramatic changes in pH in surface ocean waters that were indicated by the largest previous estimates²⁰.

There is no general agreement on the mechanism for forming rings in multi-ring basins, and a wide variety of models have been proposed^{13,23}. Prominent among these is the ring tectonic model in which material in a low-viscosity layer flows toward the crater centre as the transient cavity collapses, producing an inward drag on the base of the overlying lithosphere that creates concentric extensional faults²⁴. To form a ring by this mechanism, the transient cavity must penetrate a low-viscosity channel which, on the Earth, is most likely to be formed by the lower-continental crust²⁵ or the asthenosphere. On our new seismic data, we observe dipping reflections on all the profiles that appear to be linear faults or shear zones in the deep crust. On Chicx-C, this dipping fault-zone is directly linked to the outer ring. We therefore have direct evidence in the subsurface for the deformation associated with the formation of a ring in a multi-ring basin. If the linear fault observed on Chicx-C continued deep into the mantle, it would reach ~70 km depth beneath the centre of the crater. If the ring tectonic model is correct, our data seem to show that the required low-viscosity zone does not correspond to either the conventionally defined asthenosphere or the lower crust.

Unlike faulting in conventional tectonics, the faults we observe at Chicxulub (G and H) would have generated their entire throw in a single period of movement lasting not more than a few minutes. Such faulting can generate large volumes of melt within the fault zone to produce extensive pseudotacholytes²⁶. These are probably the reason that the faults are visible on the seismic data. Rings of pseudotacholyte are observed at the Sudbury crater, where they have been interpreted as having a fault origin, and as being associated with ring formation in a multi-ring basin²⁷. We suspect that the dipping reflectors that we observe in the seismic data are analogous to the rings of pseudotacholyte that exist at Sudbury.

The Chicxulub seismic survey has produced the first high-resolution whole-crustal structural image of a well-preserved large crater, together with direct evidence for a multi-ring basin on the Earth. Continued processing of the data will further refine these images, and particularly should improve the definition of shallow features at large radial distances. Analysis of the accompanying wide-angle data will provide additional lithological and structural constraints, especially in the deep crust, central uplift and peak ring. □

Received 18 July; accepted 14 October 1997.

- Alvarez, L. W., Alvarez, W., Azaro, F. & Michel, H. V. Extraterrestrial cause for the Cretaceous–Tertiary extinction. *Science* **208**, 1095–1108 (1980).
- Hildebrand, A. R. *et al.* A possible Cretaceous–Tertiary boundary impact crater on the Yucatan peninsula, Mexico. *Geology* **19**, 867–871 (1991).
- Swisher, C. C. III *et al.* Coeval ⁴⁰Ar/³⁹Ar ages of 65.0 million years ago from Chicxulub crater melt rock and Cretaceous–Tertiary boundary tektites. *Science* **257**, 954–958 (1992).
- Sharpton, V. L. *et al.* Chicxulub multi-ring impact basin: Size and other characteristics derived from gravity analysis. *Science* **261**, 1564–1567 (1993).
- Pope, K. O., Ocampo, A. C. & Duller, C. E. Surficial geology of the Chicxulub impact crater, Yucatan, Mexico. *Earth Moon Planets* **63**, 93–104 (1993).
- Camargo-Zanoguera, A. & Suarez-Reynoso, G. Evidencia Sismica del crater impacto de Chicxulub, G. Bol. Asoc. Mex. Geof. Expl. **34**, 1–28 (1994).
- Ward, W. C., Keller, G., Stinnesbeck, W. & Adatte, T. Yucatan subsurface stratigraphy: Implications and constraints for the Chicxulub impact. *Geology* **23**, 873–876 (1995).
- Pope, K. O., Ocampo, A. C., Kinsland, G. L. & Smith, R. Surface expression of the Chicxulub crater. *Geology* **24**, 527–530 (1996).
- Hildebrand, A. R. *et al.* Size and structure of the Chicxulub crater revealed by horizontal gravity gradients. *Nature* **376**, 415–417 (1995).
- Sharpton, V. L. *et al.* Model of the Chicxulub impact basin. *Geol. Soc. Am. Spec. Pap.* **307**, 55–74 (1996).
- Melosh, H. J. *Impact Cratering: A Geologic Process* (Oxford Univ. Press, New York, 1989).
- Alexopoulos, J. S. & McKinnon, W. B. Large impact craters and basins on Venus, with implications for ring mechanics on the terrestrial planets. *Geol. Soc. Am. Spec. Pap.* **293**, 29–50 (1994).
- Urrutia-Fucugauchi, J., Marin, L. & Trejo-Garcia, A. UNAM Scientific drilling program of the Chicxulub impact structure—Evidence for a 300-kilometre crater diameter. *Geophys. Res. Lett.* **23**, 1565–1568 (1996).
- Warner, M. R. Basalts, water or shear zones in the lower continental crust? *Tectonophysics* **173**, 163–174 (1990).
- Croft, S. K. Cratering flow fields: Implications of the excavation and transient expansion stages of crater formation. *Proc. 11th Lunar Planet. Sci. Conf.* 2347–2378 (Pergamon, New York, 1980).
- Hörz, F., Ostertag, R. & Rainey, D. A. Bunte Breccia of the Ries: Continuous deposits of large craters. *Rev. Geophys. Space Phys.* **21**, 1667–1725 (1983).
- Holsapple, K. A. & Schmidt, R. M. On the scaling of crater dimensions 2. Impact processes. *J. Geophys. Res.* **87**, 1849–1870 (1982).
- D'Hondt, S. *et al.* Surface-water acidification and extinction at the Cretaceous–Tertiary boundary. *Geology* **22**, 983–986 (1994).
- Pope, K. O., Baines, K. H., Ocampo, A. C. & Ivanov, B. A. Impact winter and the Cretaceous/Tertiary extinctions: Results of a Chicxulub asteroid impact model. *Earth Planet. Sci. Lett.* **128**, 719–725 (1994).
- Ivanov, B. A. *et al.* Degassing of sedimentary rocks due to Chicxulub impact: Hydrocode and Physical simulations. *Geol. Soc. Am. Spec. Pap.* **307**, 125–140 (1996).
- Schultz, P. H. & Merrill, R. B. *Multi-ring Basins* (Pergamon, New York, 1981).
- Melosh, H. J. & McKinnon, W. B. The mechanics of ringed basin formation. *Geophys. Res. Lett.* **5**, 985–988 (1978).
- Alexopoulos, J. S. & McKinnon, W. B. Multiringed impacts on Venus: An overview from Arecibo and Venera images, and initial Magellan data. *Icarus* **100**, 347–363 (1992).
- Spray, J. G. Super faults. *Geology* **25**, 579–582 (1997).
- Spray, J. G. & Thompson, L. M. Friction melt distribution in a multi-ring impact basin. *Nature* **373**, 130–132 (1996).

Acknowledgements. The reflection seismic data were acquired by BIRPS and funded by the Natural Environment Research Council and the BIRPS Industrial Associates programme. The project also received funding from the National Science Foundation, the Leverhulme Trust, the Royal Society, and the Royal Commission for the Exhibition of 1851. The data were acquired by Geco-Prakla and processed by Bedford Interactive Processing Services. We thank Petroleos Mexicanos for releasing their seismic data to us.

Correspondence should be addressed to M.W. (e-mail: m.warner@ic.ac.uk).

A conserved RNA-binding protein that regulates sexual fates in the *C. elegans* hermaphrodite germ line

Beilin Zhang^{*†}, Maria Gallegos^{*†}, Alessandro Puoti^{*}, Eileen Durkin^{*}, Stanley Fields[‡], Judith Kimble^{*§} & Marvin P. Wickens^{*}

^{*} Department of Biochemistry, University of Wisconsin–Madison, Wisconsin 53706, USA

[‡] Howard Hughes Medical Institute, Departments of Genetics and Medicine, University of Washington–Seattle, USA 98195

[§] Howard Hughes Medical Institute, Department of Medical Genetics and Laboratory of Cell and Molecular Biology, University of Wisconsin–Madison, Wisconsin 53706, USA

[†] These authors contributed equally to this work.

The nematode *Caenorhabditis elegans* has two sexes, males and hermaphrodites. Hermaphrodites initially produce sperm but switch to producing oocytes. This switch appears to be controlled by the 3' untranslated region of *fem-3* messenger RNA. We have now identified a binding factor (FBF) which is a cytoplasmic protein that binds specifically to the regulatory region of *fem-3* 3' UTR and mediates the sperm/oocyte switch. The RNA-binding domain of FBF consists of a stretch of eight tandem repeats and two short flanking regions. This structural element is conserved in several proteins including *Drosophila* Pumilio, a regulatory protein that controls pattern formation in the fly by binding to a 3' UTR. We propose that FBF and Pumilio are members of a widespread family of sequence-specific RNA-binding proteins.

Messenger RNAs are regulated to specify when, where and how much protein they produce by regulatory proteins that bind to specific target sites on the RNA. In much the same way, proteins bind to specific sites on DNA to control gene transcription. Regulatory sites in the 3' untranslated region (UTR) are critical for control of translation, RNA stability and localization^{1–6}. Examples are particularly prominent during early development. In *C. elegans*, 3'UTR control elements influence blastomere identity, germline cell fates and control of the life cycle⁷. In *Drosophila*, 3'UTR-based controls direct formation of the anterior–posterior axis^{6,8–11}; in *Xenopus* and mouse, they govern progression through the meiotic cell cycle^{1,12}. The RNA elements that mediate 3'UTR regulation typically do not possess obvious secondary structure, unlike many of the well characterized protein-recognition sites in RNA. Despite the importance and prevalence of these 3'UTR regulatory elements, few of their *trans*-acting factors have been identified.

In *C. elegans* hermaphrodites, specification of germ cells as sperm or oocyte is regulated by post-transcriptional controls residing in the 3'UTRs of two key sex-determining genes. Each adult germline tube contains sperm proximally and oocytes distally (Fig. 1a). The onset of spermatogenesis is governed by the 3'UTR of the *tra-2* gene¹³, whereas the switch from spermatogenesis to oogenesis is regulated by the 3'UTR of *fem-3* (ref. 14). This work focuses on the *fem-3* regulation and on the control of the sperm/oocyte switch. Loss-of-function *fem-3* mutants, *fem-3(lf)*, do not make sperm, but instead produce only oocytes (Fig. 1b, top)^{15,16}. By contrast, gain-of-function *fem-3* mutants, *fem-3(gf)*, only make sperm and do not switch to oogenesis (Fig. 1b, bottom). Therefore, *fem-3* is repressed to achieve the switch from spermatogenesis to oogenesis.

The sperm/oocyte switch appears to be controlled by a regulator that acts through a specific binding site in the *fem-3* 3'UTR. First, all *fem-3(gf)* mutations map to the 3' untranslated region: 17 of 19 mutations are single-nucleotide changes that map to a five-nucleotide stretch in the centre of the 3'UTR; this small region identifies the point mutation element, or PME (Fig. 1c)^{14,17}. Second, RNAs containing the PME bind specifically to a factor present in worm extracts¹⁸. Third, overexpression of an RNA harbouring the

wild-type *fem-3* 3'UTR masculinizes the nematode germ line, consistent with titration of a repressor from the wild-type *fem-3* gene¹⁴. Therefore, the 3'UTR of *fem-3* appears to interact with a specific RNA-binding factor to achieve post-transcriptional regulation.

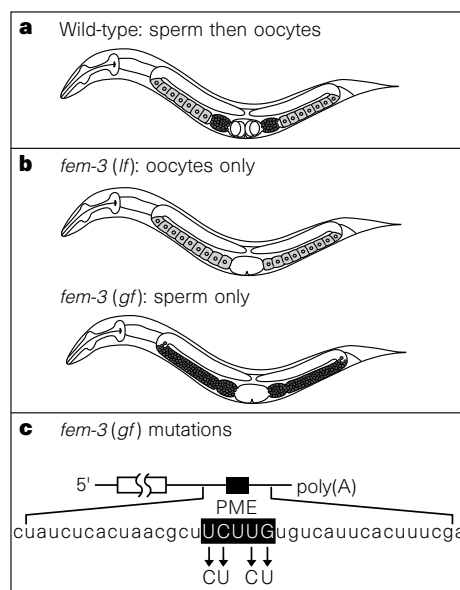


Figure 1 Phenotypes and gain-of-function alleles of *fem-3*. **a**, Wild-type hermaphrodite adult. Anterior is left, dorsal is up. The gonad consists of two arms that share a central uterus. Within each gonadal arm, sperm (dark grey) are proximal and oocytes (light grey) distal. **b**, Adult *fem-3* mutant phenotypes. Dark grey, sperm; light grey, oocytes. **c**, Molecular basis of *fem-3(gf)* mutations. Sequence of the 37 nt of the *fem-3* 3'UTR used for three-hybrid screening; black box, the PME^{14,18}. *fem-3(gf)* point mutations are depicted. The most severe *fem-3(gf)* point mutation, *fem-3(q96)*, changes G to T at the 3' end of the PME.

Here we use a yeast three-hybrid assay¹⁹ to identify a protein, designated FBF (for *fem-3* binding factor), that binds specifically to the PME in the *fem-3* 3'UTR. We show that FBF is required for regulation of the pattern of sexual fates in the hermaphrodite germ line. FBF is structurally related to Pumilio, which regulates pattern formation in the early *Drosophila* embryo²⁰. We propose that FBF and Pumilio are members of a widespread family of RNA-binding proteins.

Identification of FBF

To identify a protein that interacts specifically with the *fem-3* PME, we employed a modified form of the yeast three-hybrid system (Fig. 2a)¹⁹. In this assay, interaction of an RNA-binding protein with its cognate RNA binding site leads to transcriptional activation of a reporter gene in yeast.

We first introduced a plasmid encoding a bifunctional ('hybrid') RNA into a yeast strain expressing a LexA-MS2 coat protein fusion¹⁹. That hybrid RNA possessed a tandem duplication of 37 nucleotides of the *fem-3* 3'UTR, which includes the PME, and MS2 coat protein recognition sites (Fig. 2b)¹⁹. The multicopy plasmid encoding the hybrid RNA carried both *URA3* and *ADE2* markers. The MS2 sites in the hybrid RNA interact with the MS2 coat protein portion of a LexA-MS2 coat protein fusion, which in turn binds to LexA operators positioned upstream of both *HIS3* and *lacZ* genes (Fig. 2a)¹⁹.

Into a strain harbouring the PME-MS2 hybrid RNA plasmid, we introduced a library of *C. elegans* complementary DNAs linked to the yeast *GAL4* transcriptional activation domain (a gift from R. Barstead). Cells were selected for increased *HIS3* expression, but not for retention of the hybrid RNA plasmid. This protocol yielded two classes of *HIS3*⁺ transformants: 'hybrid RNA-dependent' ones that require the hybrid RNA plasmid for survival, and 'hybrid RNA-

independent' ones that do not. To identify hybrid RNA-dependent candidates, we exploited the *ADE2* marker on the hybrid RNA plasmid: cells lacking this plasmid are red owing to a lack of the *ADE2*-encoded protein, whereas those carrying the hybrid plasmid are white. White candidates that activated both *HIS3* and *lacZ* genes were next treated with 5-fluoro-orotic acid to select against the *URA3* marker and thereby eliminate the RNA plasmid. Candidates that failed to activate *lacZ* following plasmid loss were next tested for PME binding specificity by reintroducing hybrid RNA plasmids carrying either wild-type or a mutant form of the *fem-3* 3'UTR. A single cDNA/activation domain plasmid, pBZ1, satisfied all screening criteria.

The RNA-binding specificity of the protein encoded by pBZ1 was assayed more stringently using a battery of hybrid RNA plasmids (Fig. 2c). *LacZ* expression was activated in the presence of a hybrid RNA containing the wild-type *fem-3* PME (Fig. 2c, row 1), but not with a similar RNA bearing either an eight-nucleotide mutant centred on the PME (row 2) or a single nucleotide change in the PME (row 3). This latter mutation corresponds to *fem-3*(q96), the phenotypically strongest *fem-3*(gf) point mutation¹⁴. Furthermore, *lacZ* was not activated by hybrid RNAs in which the *fem-3* 3'UTR segment was replaced with an iron response element (IRE, row 4), an oligo(A)₃₀ tract (row 5), or a 574-nucleotide segment of HIV RNA (row 6). The amount of wild-type (WT) *fem-3* hybrid RNA was not significantly greater than that of the control RNAs, as judged by northern blotting (results not shown). We conclude that the protein encoded by pBZ1, which we call FBF-1, binds specifically to the wild-type PME.

The *C. elegans* portion of pBZ1 was used to search the *C. elegans* genomic sequence²¹ and to screen cDNA libraries. Two almost identical cDNAs were identified that are encoded by two distinct genes, *fbf-1* and *fbf-2* (Fig. 2d). These two genes are located on

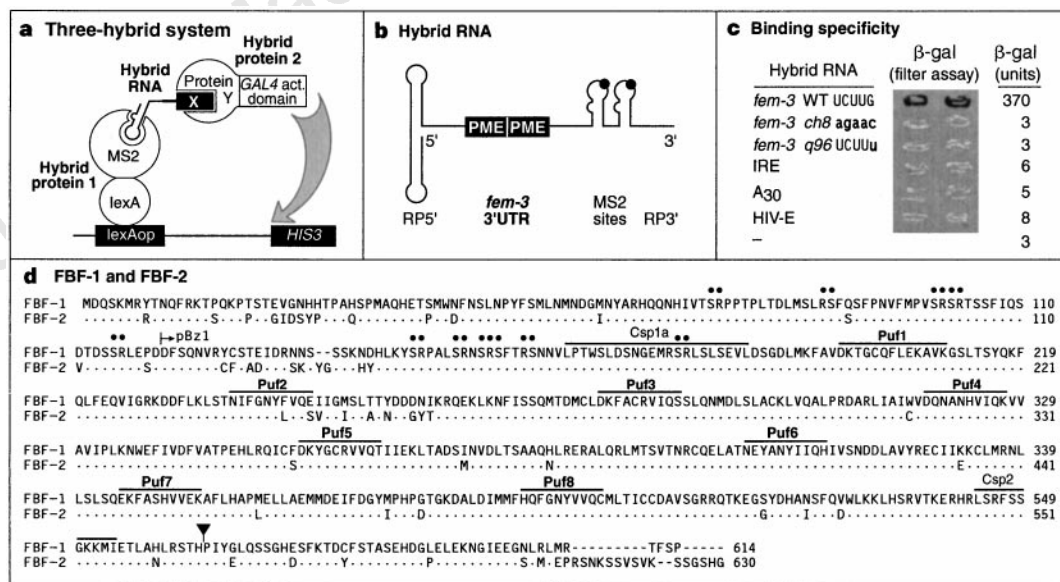


Figure 2 Isolation and sequence of FBF-1 and FBF-2. **a**, The yeast three-hybrid system. Hybrid protein 1 comprises the LexA DNA-binding protein and MS2 coat protein. Hybrid protein 2 comprises an RNA-binding domain, designated protein Y, linked to the yeast *GAL4* transcriptional activation domain. The hybrid RNA, which links the two-hybrid proteins, comprises a binding site for MS2 coat protein linked to the RNA sequence of interest, designated RNA-X (box). **b**, The hybrid RNA. From 5' to 3', the hybrid RNA consists of the 5' leader of RNase P RNA (RP5'), two tandem repeats of a segment of the *fem-3* 3'UTR containing the PME (rectangles; sequence in Fig. 1c), two binding sites for MS2 coat protein including a point mutation (black dot) that increases binding affinity, and the 3' trailer sequence of RNase P RNA (RP3'). **c**, Binding specificity. pBZ1 was tested for RNA

binding specificity with several different hybrid RNAs. '*fem-3* WT UCUUG' contains the 37-nt segment of wild-type *fem-3* 3'UTR shown in Fig. 1c; '*fem-3* ch8 agaac' and '*fem-3* q96 UCUUu' carry mutant versions of the segment (see text). 'IRE', 'A30' and 'HIV-E' contain the indicated RNA sequences in place of *fem-3* sequences (see Methods). Filter and liquid assays are shown. **d**, Predicted FBF-1 and FBF-2 sequences. In FBF-2, only amino acids that diverge from the FBF-1 sequence are shown. Dots, identical amino acids; dashes, missing amino acids. Lines above the sequence indicate the core consensus sequence of each Puf repeat, and the conserved regions, Csp1a and Csp2. Dots, SR or RS dipeptides; solid triangles, ends of FBF-1Δ13 (Fig. 6b); line with arrow, region of FBF-1 present in pBZ1 (amino acids 121 to 614).

chromosome II near *lin-4* within a region represented by a single 450-kb YAC (Y56A9), but are not immediately adjacent in the chromosome (see Methods). The complete coding regions of *fbf-1* and *fbf-2* are 93% identical in nucleotide sequence, and 91% identical in amino-acid sequence. pBZ1 encodes the C-terminal 81% of FBF-1 (corresponding to amino acids 121 to 614; Fig. 2d). FBF-2 displays the same RNA-binding specificity as FBF-1 in the three-hybrid assay, and no differences between *fbf-1* and *fbf-2* have been detected in assays of *fbf* function or expression (see below). Therefore, *fbf-1* and *fbf-2* may be redundant (see Regulation of sperm/oocyte switch by FBF, below). From now on, we refer to *fbf-1* and *fbf-2* genes collectively as *fbf* and to the protein as FBF.

Germline-specific *fbf* mRNA and protein

The binding specificity of FBF to the *fem-3* PME makes it a good candidate for the repressor that is required for the sperm/oocyte switch. We therefore asked whether *fbf* would be expressed appropriately during development. To examine *fbf* mRNA, we analysed polyadenylated *C. elegans* RNA on a northern blot using the insert as a probe. A major mRNA of 2.4 kilobases (kb) was observed (Fig. 3a). As this probe detects both *fbf-1* and *fbf-2* genes in a Southern blot (result not shown), we infer that it detects both mRNAs. *fbf* mRNA increases in abundance during post-embryonic development and peaks during the fourth larval stage, when the sperm/oocyte switch occurs (L4; Fig. 3a, lanes 1–5). *fbf* mRNA is not detected in *glp-1* mutants, which lack a germ line (Fig. 3a, lane 6)²². The simplest interpretation of these results is that *fbf* mRNA is restricted to the germ line.

To examine FBF protein, we prepared affinity-purified polyclonal antibodies against a bacterially expressed, hexahistidine-tagged portion of FBF-1 (amino acids 121–614, Fig. 2d). Studying whole nematodes by immunofluorescence, we detected FBF only in the

germ line and found that it is predominantly cytoplasmic (Fig. 3b). Staining with anti-DNA antibodies reveals nuclei in all tissues of the same animals (not shown). Because the antibodies were raised against a segment of the protein that is 92% identical at the amino-acid level in FBF-1 and FBF-2, they are likely to detect both proteins. Consistent with this, antibody staining was eliminated following injection of either *fbf-1* or *fbf-2* antisense RNA (see below). Localization of FBF to the germline cytoplasm is consistent with a role in post-transcriptional regulation of *fem-3* mRNA to achieve the sperm/oocyte switch.

Sperm/oocyte switch requires *fbf*

Removal of a repressor is predicted to have the same phenotypic effect as elimination of the regulatory element through which it acts. Because disruption of the PME in *fem-3(gf)* mutants leads to a failure in the switch from spermatogenesis to oogenesis, loss of *fbf* might likewise lead to an excess of sperm and prevent oogenesis. To examine the phenotype of animals lacking *fbf*, we used a reverse genetic assay termed RNA-mediated interference (RNAi). In this technique, RNA corresponding to the coding region of a given gene is injected into the *C. elegans* germ line, causing a lack of expression from the same gene in the injected animal's progeny. This procedure reproduces the phenotypes associated with loss-of-function mutations in most of the genes tested so far (see refs 23, 24, for example). The mechanism by which RNAi interferes with gene expression appears not to be simple inhibition via base pairing of the injected and target RNAs, because either antisense or sense RNA can be used. We injected antisense RNA to avoid artefacts that might result from titration of factors binding to sense RNA¹⁴. We refer to progeny of hermaphrodites injected with *fbf* RNA as *fbf(RNAi)* animals.

We first determined the effect of *fbf(RNAi)* on the abundance of FBF protein. Antisense RNA was synthesized *in vitro* either from the *fbf* coding region of pBZ1 or, as a control, from a second *C. elegans* gene, F54C9.8, which is related to *fbf* (Fig. 6c). The germ lines of *fbf-1(RNAi)* animals showed greatly reduced staining by anti-FBF antibodies (Fig. 3b), as did those of *fbf-2(RNAi)* animals (not shown). By contrast, injection of *fbf* antisense RNAs did not prevent staining by anti-DNA or anti-tubulin antibodies (not shown). We conclude that injection of *fbf* antisense RNA dramatically and specifically reduces FBF protein levels, and so provides a tool for investigating *fbf* function.

We next examined the phenotypic effect of *fbf(RNAi)*. By Nomarski microscopy, most *fbf(RNAi)* animals produced excess sperm and no oocytes (Fig. 4a, left), similar to the *fem-3(gf)* phenotype. By contrast, control F54C9.8(*RNAi*) animals showed no detectable effect, and yielded animals with a typical number of sperm and oocytes (Fig. 4a, right). As an additional test for gamete differentiation, we stained both *fbf(RNAi)* and control F54C9.8(*RNAi*) animals using anti-sperm antibodies²⁵ and DAPI staining. As expected, *fbf(RNAi)* animals showed a vast increase in both the number of their sperm and antibody staining (Fig. 4b). DAPI staining revealed highly condensed nuclei, typical of sperm, over a much larger region of the gonad in *fbf(RNAi)* than in control germ lines (Fig. 4b). Quantitatively, 95% of *fbf(RNAi)* animals produced sperm only, whereas no control F54C9.8(*RNAi*) animals had this phenotype (Fig. 4c). In addition to germline masculinization *fbf(RNAi)* animals often had small germ lines and made abnormal oocytes, suggesting that *fbf* may also play a role in germline proliferation and oogenesis. However, no other defects (such as sexual transformation of somatic cells, embryonic lethality, uncoordinated behaviour or morphological defects) were observed in either *fbf(RNAi)* or F54C9.8(*RNAi*) animals.

We conclude that loss of *fbf* results in a dramatic masculinization of the germ line. This phenotype strongly supports the hypothesis that *fbf* is required for the switch from spermatogenesis to oogenesis.

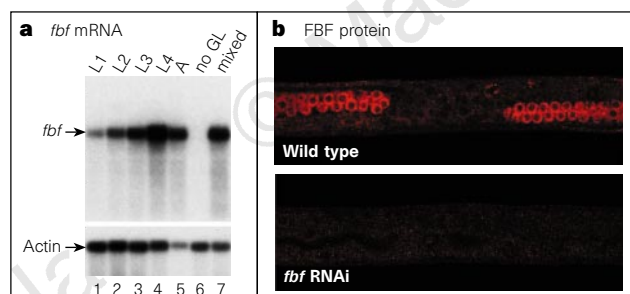


Figure 3 *fbf* mRNA and protein. **a**, Northern blot of *fbf* mRNA. Polyadenylated *C. elegans* mRNA was separated through a 1% agarose-glyoxal gel, transferred to a filter, and probed with *fbf-1* cDNA (nt 489 to 1,995). Lanes 1–5, RNA from staged animals: L1–L4, four larval stages; A, adults. Lane 6, no germ line (GL): RNA was prepared from *glp-1(q224)* adults raised at 25 °C, which contain no more than 8 germ cells, as compared to more than 2,000 in wild type²². Lane 7, mixed stages: RNA was derived from L1-to-adult stage animals. Arrow for *fbf* indicates a 2.4-kb band corresponding to *fbf-1* and *fbf-2* transcripts; arrow for 'actin' points to a band derived by probing the same blot for actin mRNA as standard. **b**, FBF protein detected by immunofluorescence in whole-mount animals. Central views of each animal are shown and include all of the developing germ line, which is proliferating at this stage of development. Top, wild-type L3 hermaphrodite. Germ line nuclei are dark circles that are surrounded by intensely stained cytoplasm. Bottom, *fbf-1(RNAi)* L3 hermaphrodite. Staining is eliminated.

fbf in the sex-determination pathway

The sex-determining genes of *C. elegans* have been ordered in a regulatory pathway, largely by analysis of double mutants (reviewed in ref. 26). Included in this pathway are genes that act in all tissues to regulate male or female states (for example, *fem-3*), and genes that act in a single tissue to regulate specific sexual fates (such as *fog-2* (for feminization of the germ line)). We sought to place *fbf* in this pathway. Specifically, we determined whether *fbf* acted in a position consistent with its proposed role as a negative regulator of *fem-3*. At the same time, we compared the effect of loss of *fbf* to that of a *fem-3(gf)* mutation in various genetic backgrounds; we reasoned that if *fbf* does act through the *fem-3* 3'UTR, loss of *fbf* and disruption of the *fem-3* PME should yield similar results.

We first performed an epistasis test using *fbf-1(RNAi)* and a *fem-3* null mutant (Fig. 5a, rows 1–3). If germline masculinization of *fbf(RNAi)* animals results from a failure to repress *fem-3*, then this phenotype should require wild-type *fem-3*. Normally, *fem-3* null mutants make only oocytes and produce no self-progeny (row 1)^{15,16}. To test whether *fem-3* mutants could be masculinized by a reduction of *fbf*, we used a *fem-3* null mutation marked with a closely linked marker, *unc-24(lf)*, which exhibits an uncoordinated (Unc) phenotype. After injection of *fbf* antisense RNA into a *unc-24(lf) fem-3(lf)* homozygote, animals were fertilized by heterozygous *unc-24(lf) fem-3(lf)/++* males. Progeny from the cross include non-Unc heterozygotes of genotype *unc-24(lf) fem-3(lf)/++* (row 2), and Unc homozygotes of genotype *unc-24(lf) fem-3(lf)* (row 3). We found that 100% of the non-Unc heterozygous progeny made only sperm (row 2), confirming the efficiency of RNAi in this

experiment. By contrast, Unc siblings, which lack both maternal and zygotic *fem-3*, made no sperm; instead, 69% made only oocytes (row 3). The remaining animals produced neither sperm nor oocytes. These results show that the *fbf(RNAi)* phenotype requires wild-type *fem-3*, as expected if *fbf* is a *fem-3* repressor.

To further position *fbf* in the sex determination pathway, we injected *fbf-1* antisense RNA into *fog-2(lf)* or *her-1(lf)* homozygotes (Fig. 5a, rows 4–10). Both *fog-2* and *her-1* are normally required for male development, *fog-2* specifically for sperm and *her-1* more generally for male development. Both *fog-2* and *her-1* are predicted to act upstream of *fem-3* (refs 27, 28). Normally, XX *fog-2* mutants produce only oocytes (row 4)²⁷. By contrast, either *fog-2; fem-3(gf)* double mutants or *fog-2;fbf-1(RNAi)* animals are masculinized (rows 5 and 6)²⁷. Similarly, XO *her-1* mutants produce both sperm and oocytes (row 8)²⁸, but can be masculinized either by *fem-3(gf)* (row 9)¹⁶ or by *fbf(RNAi)* (row 10).

Taken together, these results indicate that loss of *fbf* and disruption of the PME yield similar phenotypes. We conclude that *fbf* lies upstream of *fem-3*, but downstream or parallel to *her-1* and *fog-2*, consistent with its proposed role as a negative regulator of *fem-3* (Fig. 5b).

FBF structure and RNA-binding domain

The most striking feature of the amino-acid sequences of FBF-1 and FBF-2 is the presence of eight repeats with a highly conserved core consensus sequence (Fig. 6a). Within the repeat region, FBF-1 and FBF-2 are nearly identical, deviating in only a single amino acid among the eight core consensus sequences. In addition, 13 and 14

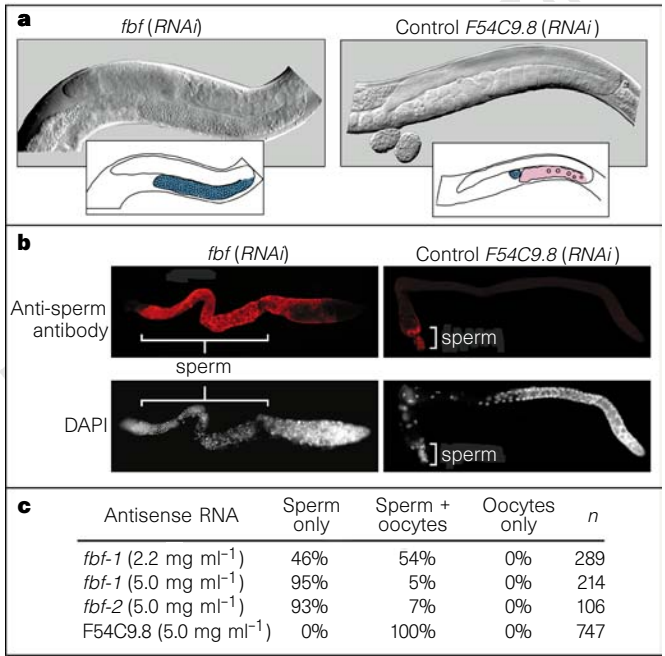


Figure 4 Tests of *fbf* function. **a**, *fbf(RNAi)* phenotype by Nomarski. Top, Nomarski differential interference contrast micrograph; bottom, drawing of the same animal, highlighting locations of sperm (blue) and oocytes (pink). Left, *fbf(RNAi)* adult. Sperm are produced in excess; no oocytes are observed. Right, control F54C9.8(RNAi) adult. Sperm and oocytes are made in normal amounts. **b**, *fbf(RNAi)* phenotype shown by fluorescence microscopy. Extruded adult gonads were stained with either anti-sperm antibody (top) or DAPI (bottom). In DAPI staining, small dots are highly condensed sperm nuclei; brackets indicate the region containing such nuclei. Left, *fbf(RNAi)* gonads; right, F54C9.8(RNAi) gonads. **c**, Quantification of the effects of *fbf(RNAi)*. Left column shows the type and concentration of antisense RNA injected; *n* is the number of gonadal arms analysed.

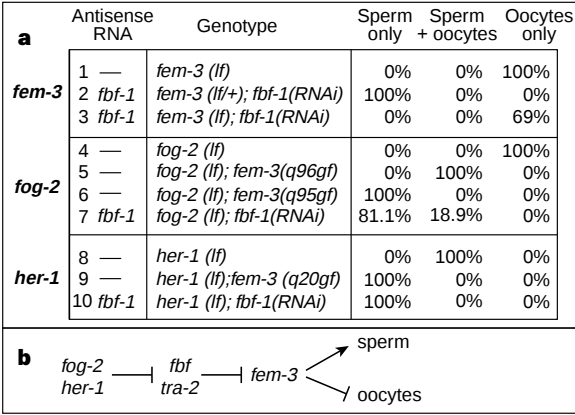


Figure 5 Positioning *fbf* in the sex determination pathway. **a**, Results of epistasis experiments. *fem-3* (rows 1–3), row 1, *fem-3* null mutant makes only oocytes¹⁵. Row 2, *fem-3(e1996/+)* heterozygotes treated with *fbf* RNAi (*n* = 49 gonadal arms). Row 3, *fem-3(e1996)* homozygotes treated with *fbf* RNAi (*n* = 41 gonadal arms). Among animals that initiated oogenesis, 62% of the gonadal arms produced oocytes of wild-type appearance. Among progeny with neither sperm nor oocytes, germ cells were apparently undifferentiated and difficult to classify. *fog-2* (rows 4–7), phenotypes of *fog-2* single mutant (row 4) or *fog-2; fem-3(gf)* double mutants (rows 5 and 6)²⁷. *fem-3(gf)* alleles masculinize *fog-2(lf)* alleles to different degrees, depending on the strength of the *fem-3(gf)* allele used; *fem-3(q95gf)*, a deletion of the PME region^{14,16}, is stronger than *fem-3(q96gf)*¹⁶. Row 7, *fog-2* homozygotes that were treated with *fbf* RNAi (*n* = 111 gonadal arms). *her-1* (rows 8–10), phenotypes of *her-1(lf)* single mutant (row 8) or *her-1; fem-3(gf)* double mutants (rows 9)²⁷. Row 10, *her-1* homozygotes treated with *fbf* RNAi (*n* = 41 gonadal arms). **b**, Position of *fbf* in sex determination hierarchy. The epistasis analysis above suggests that *fbf* acts downstream of *fog-2* and *her-1*, but upstream of *fem-3*, in the previously determined sex determination pathway. This is the same epistasis level as *tra-2*, another predicted negative regulator of *fem-3* (ref. 15). Only genes discussed in the text are depicted.

SR dipeptides are found within the N-terminal regions of FBF-1 and FBF-2 respectively, suggesting that these proteins may contain an SR domain (Figs 2d, 6a). SR domains are found in several general and alternative splicing factors, and may mediate protein–protein interactions^{29,30}.

The repeats of FBF are similar to those of *Drosophila* Pumilio (Fig. 6a)^{31,32}. Pumilio binds to *nanos*-response elements (NREs) in the 3'UTR of *hunchback* (*hb*) mRNA, thereby causing translational repression and destabilization of *hb* mRNA^{6,20}. Both FBF and Pumilio contain eight repeats with a similar conserved core consensus sequence (Fig. 6a)³². Furthermore, comparison of individual repeats in FBF-1 and Pumilio reveals a conserved pattern from repeat to repeat (see below). Although the FBF and Pumilio repeats are similar, they differ in two ways. The spacings between adjacent core consensus sequences in FBF are more heterogeneous than in Pumilio (Fig. 6a) and most conserved amino acids in the FBF repeats lie in the core consensus sequence, whereas those in Pumilio extend more broadly through the repeat. Thus FBF and Pumilio appear to be divergent members of a protein family, as discussed later.

To identify the region of FBF required for RNA binding, we analysed a series of FBF-1 deletion mutants using the yeast three-hybrid system (Fig. 6b). Deletion of the N-terminal region containing most of the putative SR domain did not alter the interaction, as monitored by the level of β -galactosidase (Fig. 6b, D1 to D3). D3, which still activates *lacZ* expression, leaves only 25 amino acids before the start of the first repeat. However, deletion of an additional 20 amino acids eliminated *lacZ* activation (Fig. 6b, D4). More extensive N-terminal deletions did not activate the reporter gene either (Fig. 6b, D5 and D11). Deletion derivatives lacking 28 or 47 amino acids from the C terminus (D6 and D7) stimulated *lacZ* expression, but removal of an additional 20 residues abolished activation entirely (Fig. 6b, D8). Further C-terminal truncations also did not activate (Fig. 6b, D9 and D10). As judged by western blotting of yeast extracts, protein derivatives that fail to activate (D4, 5, 8, 9 and 10) are present at levels comparable to that of the wild-type protein (not shown).

To define the minimum contiguous portion of FBF-1 that can interact with *fem-3* RNA, N- and C-terminal deletions were introduced into the same constructs (Fig. 6b, D12 to D14). An FBF-1 derivative that extends from 25 amino acids before the first repeat to 38 amino acids after the last, activates transcription of *lacZ* (Fig. 6b, D13). Within the region present in D13, *fbf-1* and *fbf-2* are 95% identical (Fig. 2d). Together, our findings suggest that the eight repeats plus short flanking regions on each side may be required for the binding of FBF protein to *fem-3* RNA.

Regulation of sperm/oocyte switch by FBF

FBF seems to be a central component of the *fem-3* repressor that regulates the switch from spermatogenesis to oogenesis in *C. elegans* hermaphrodites. FBF has the correct RNA-binding specificity, its loss eliminates the switch in cell fate, it occupies the appropriate level in the genetic hierarchy of sex-determining genes, and it is expressed at the right time and place. We propose that repression of the *fem-3* gene requires binding of FBF to the 3'UTR and that this FBF-mediated repression leads to oogenesis (Fig. 7, left). The molecular mechanism of repression is post-transcriptional, and may include effects on translation, mRNA stability or transport.

The model shown in Fig. 7 depicts regulated binding of FBF; other models are also tenable. For example, FBF could bind the *fem-3* PME continuously throughout development, and other factors could act together with FBF to achieve the switch at the appropriate time. The six *mog* (for masculinization of the germ line) genes could encode such regulators of FBF activity, as they too are required for the sperm/oocyte switch^{33,34} and for PME-mediated regulation (M.G. and J.K., manuscript in preparation). As FBF is confined to the germ line but *fem-3(gf)* mutants can exhibit somatic

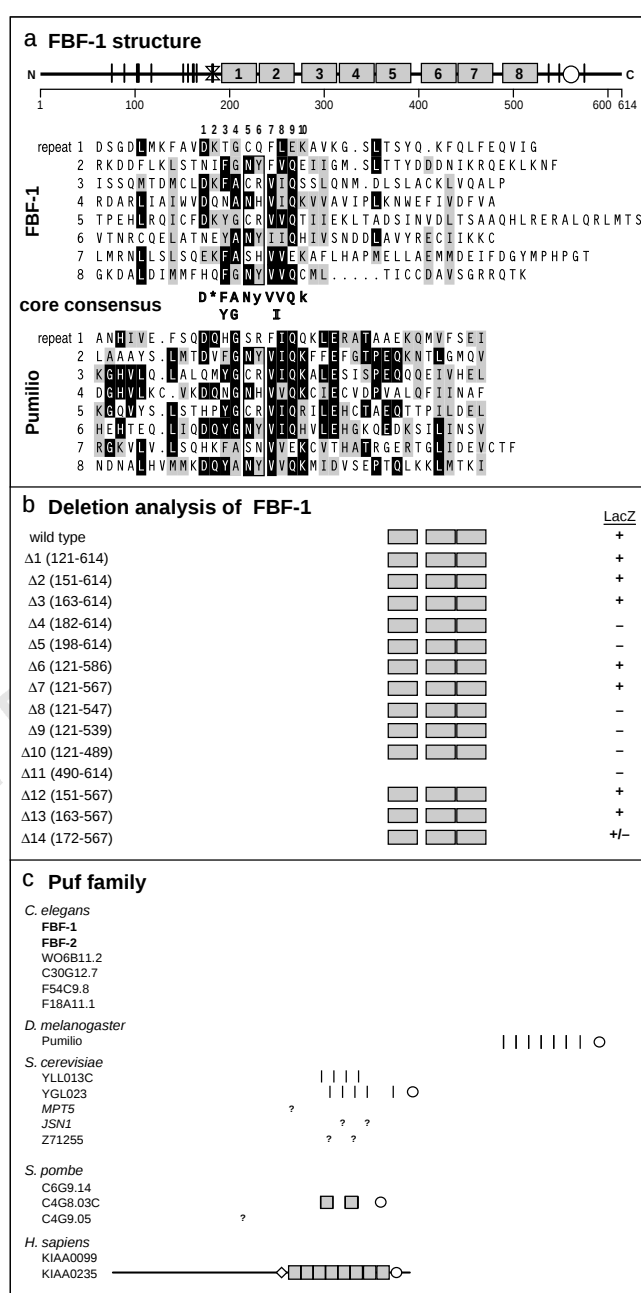


Figure 6 A family of related proteins containing multiple Puf repeats. a, Comparison of repeats in FBF-1 and Pumilio. Top, FBF-1 structure. Vertical lines, SR and RS dipeptides^{29,30,38}; anvil, Csp1a domains; grey rectangles, Puf repeats; circle, Csp2 domain. Below, continuous sequence of FBF-1 and Pumilio through the Puf repeat region, with core consensus sequences aligned. Residues in black boxes are present in four or more repeats; residues in grey boxes are either related to the most conserved residue or present in three repeats of FBF. The shared core consensus sequence (numbered above the FBF sequence) is provided: upper case, amino acids present in four or more repeats; lower case, amino acids present in three repeats of FBF. The black rectangle surrounding positions 5 and 6 of the consensus highlights the conserved pattern of amino acids from one repeat to the next. b, Deletion analysis of FBF-1 to define the minimum contiguous portion of FBF-1 required for RNA binding. Sequence motifs are depicted as in a. Deletion endpoints in parentheses, using numbering in Fig. 2d. +, Strong blue signal in *lacZ* filter assay; +/-, weak blue; -, no colour. c, Puf protein family. Sequence motifs are as in a; diamonds, Csp1b domains. See Methods for criteria used in defining each motif, for assigning proteins to the family and for accession numbers. Additional human partial cDNAs that apparently encode Puf proteins have been identified as partial cDNAs in *Arabidopsis*, rice, rats, and mice. These are not depicted.

phenotypes (ref. 35), we infer that a distinct *fem-3* repressor exists in the soma.

We have been unable to distinguish between *fbf-1* and *fbf-2* function using RNAi— injection of one or both RNAs has the same effect with respect to gamete differentiation and elimination of anti-FBF antibody staining. The simplest interpretation is that these two nearly identical genes are redundant. However, it is possible that only one encodes the true *fem-3* repressor and that RNAi does not distinguish between them for technical reasons. Extensive genetic screens for mutants with phenotypes similar to that of *fbf(RNAi)* uncovered mutations in six *mog* genes, but did not yield mutations in *fbf-1* or *fbf-2* (refs 33,34). The inability to isolate *fbf* mutants is consistent with their proposed redundancy. To test that idea, we are currently generating *fbf* mutants by reverse genetics.

The Puf family and 3'UTR control

FBF and Pumilio are members of a large and evolutionarily widespread protein family, with members in *C. elegans*, *Drosophila*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae* and humans (Fig. 6c). We refer to these family members as Puf proteins (for Pumilio and FBF). We suggest that Puf proteins comprise a family of sequence-specific RNA-binding proteins and that the eight Puf repeats plus short flanking sequences constitute the RNA-binding domain. So far, FBF and Pumilio are the only Puf proteins with known biochemical activities: both are repressors that act by binding to specific elements in the 3'UTRs of their respective RNA targets (this work, and ref. 20). We propose that other Puf proteins may have similar functions.

The existence of the family and several key features of the repeats were first suggested by a sequence comparison between Pumilio and the yeast protein YGL023 (ref. 32), and later extended to yeast *Mpt5p* (ref. 36). Because many more sequences are now available, diagnostic features of Puf family members can be established more definitively (see Methods). Most of the 17 family members shown in Fig. 6c possess eight tandem repeats of a decapeptide core consensus sequence, reiterated at 35–40-amino-acid intervals (as depicted for FBF and Pumilio in Fig. 6a). From one repeat to the next, most Puf proteins display a characteristic pattern of amino-acid changes at several positions. For example, in consecutive repeats of both FBF and Pumilio, positions 5 and 6 of the core consensus are NCNCNSN and YRHRYxY respectively (Fig. 6a, boxed). Similarly, the seventh repeats of different Puf proteins are more similar to one another than they are to other repeats in the same protein. These higher-order patterns, which are characteristic of the entire Puf protein

family (see Methods), suggest a level of structural organization that supersedes the single repeat.

Many Puf proteins also contain conserved segments before the first repeat and after the last, which we dub Csp1 and Csp2, respectively (for conserved sequences flanking Puf repeats). The Csp domains are required for RNA binding and divide the Puf family into subclasses. Immediately preceding the first Puf repeat, FBF-1, FBF-2 and F54C9.8 of *C. elegans* contain one consensus sequence (Csp1a), whereas *Drosophila* Pumilio, yeast YLL013c and human KIAA0099 and D87078 contain another (Csp1b). Both Csp1a and Csp2 are required for RNA binding by FBF: the N-terminal boundary of FBF's RNA-binding domain (between positions 163 to 182 ($\Delta 3$ and $\Delta 4$ in Fig. 6b)) coincides with the N-terminal edge of Csp1a (position 166); its C-terminal boundary (between positions 547 ($\Delta 8$) and 567 ($\Delta 7$)) brackets the C-terminal edge of Csp2 (position 556). Similarly, with Pumilio, the minimal portion of the purified protein required for binding to RNA *in vitro* includes the eight Puf repeats plus short N- and C-terminal flanking regions (ref. 37, and Y. Murata and R. Wharton, personal communication).

The modular organization of FBF and Puf proteins superficially resembles that of other classes of RNA-binding proteins, in which domains are reiterated (such as the RRM domains)^{38,39}. However, Puf proteins possess an unusual architecture in that most family members contain the same large number of repeats and exhibit a conserved pattern of residues from one repeat to the next. Although two proteins in budding yeast (*JSN1* and *Z71255*) contain fewer Puf repeats, they may oligomerize to attain the same eight-membered arrangement or possess cryptic repeats.

The similar sequences and architecture of Puf proteins raises the important issue of whether they discriminate related RNA sequences. The *cis*-acting elements required for *fbf* and Pumilio bear only modest similarity. The *fem-3* PME region includes UCuUGu, where capital letters designate nucleotides affected by *fem-3(gf)* point mutations¹⁴; the *pumilio* target sequence in *hunchback* mRNA, the NRE (Nanos response element), includes GUUGU, where capital letters depict nucleotides conserved in several NREs⁴⁰. The significance of the shared UUGU tetranucleotide is not known. Although point mutations in the NREs of *hunchback* mRNA disrupt translational regulation⁴¹, biochemical delineation of FBF and Pumilio binding sites is needed to compare their specificities. Given the size of the putative RNA-binding region of FBF, its RNA-binding site is likely to extend beyond the four nucleotides identified by the point mutations.

Table 1 Puf family proteins

Accession number	Gene name	Species	Puf repeats	Matches at position 5	Matches at position 6
–	<i>fbf-1</i>	<i>C. elegans</i>	8	7/7	6/6
U23176	<i>fbf-2</i> (F21H12.5)	<i>C. elegans</i>	8	7/7	6/6
U39854	<i>W06B11.2</i>	<i>C. elegans</i>	8	6/7	6/6
U21319	<i>C30G12.7</i>	<i>C. elegans</i>	8	6/7	6/6
Z49967	<i>F54C9.8</i>	<i>C. elegans</i>	8	6/7	5/6
Z81507	<i>F18A11.1</i>	<i>C. elegans</i>	8	6/7	5/6
L07943	<i>pumilio</i>	<i>D. melanogaster</i>	8	7/7	6/6
Z73118	<i>YLL013c</i> (orf)	<i>S. cerevisiae</i>	8	6/7	5/6
S57889	<i>YGL023</i>	<i>S. cerevisiae</i>	8	6/7	5/6
D26184	<i>MPT5</i> (<i>HTR1</i>)	<i>S. cerevisiae</i>	8	6/7	5/6
L43493	<i>JSN1</i>	<i>S. cerevisiae</i>	2–5	N/A	N/A
Z71255	<i>YPR039c</i>	<i>S. cerevisiae</i>	2–5	N/A	N/A
Z81317	<i>SPAC6G9.14</i>	<i>S. pombe</i>	8	6/7	5/6
Z56276	<i>SPAC4G8.03c</i>	<i>S. pombe</i>	8	5/7	5/6
Z69727	<i>SPAC4G9.05</i>	<i>S. pombe</i>	8	4/7	2/5
D43951	<i>KIAA0099</i>	<i>H. sapiens</i>	8	7/7	6/6
D87078	<i>KIAA0235</i>	<i>H. sapiens</i>	8	7/7	6/6

Csp1: The Csp1a consensus is L-P-x-W-x-L/V-D-x-x-G-x-M/I-R-x-x-L-S/T-L-x-x-V-L/V, and spans residues 166 to 187 of FBF-1 (Fig. 2d), FBF-2 and F54C9.8 of *C. elegans* conform precisely to this consensus. The Csp1b consensus is G-R-S-R-L-L-E-D-F-R-N-x₀₋₅-N-x-F/Y-P-N-L-Q-L-R/K-E/D-L/I, and spans residues 1,091 to 1,112 of Pumilio (numbering in refs 31, 32). The human proteins KIAA0099 and KIAA0235, which are closely related overall, match this consensus precisely; *Drosophila* Pumilio deviates at a single position, and *S. cerevisiae* YLL013C deviates at six positions. Csp2: The Csp2 consensus is L/I-R/K-K/R-F/Y-x-x-G-K/R-K/R/H-I-I/L, and spans residues 547 to 557 of FBF-1 (Fig. 2d). To be classified as containing a Csp2 element in Fig. 6c, a protein must have a sequence at the appropriate location downstream of the final Puf repeat that conforms to at least seven of the eleven criteria imposed by this consensus (nine amino-acid identities and a two-amino-acid spacing). The precise boundaries between Csp2 and the last Puf repeat are arbitrary as entire regions from the Csp to the Puf repeat are conserved.

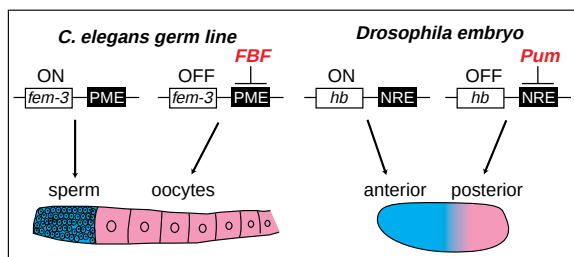


Figure 7 Comparison of regulation by FBF and Pumilio. Left, in the *C. elegans* gonad, expression of *fem-3* causes germ cells at one end of the tubular gonad to enter spermatogenesis, whereas repression of *fem-3* by FBF causes germ cells in an adjacent block of cells to enter oogenesis. Right, in the *Drosophila* embryo, expression of *hunchback* leads to more anterior fates, whereas repression of *hunchback* RNA by Pumilio leads to more posterior fates. The differences in pattern due to FBF are achieved through temporal regulation of *fem-3*; those due to Pumilio are achieved through spatial control by a localized regulator, Nanos (see text).

The isolation of *fbf-1* relied on the use of a modified form of the yeast three-hybrid system¹⁹. The *fem-3* PME region is not predicted to have extensive secondary structure. The success of our search suggests that other RNA regulatory elements that also lack predicted secondary structure, as have been found in many 3'UTRs, may be amenable to analysis by this three-hybrid system, and that the existence of nonspecific RNA-binding proteins in yeast and in cDNA libraries are not prohibitive problems. The use of a colony colour assay to eliminate RNA-independent positives, as reported here, simplifies identification of the rare desired transformants.

FBF and Pumilio in development and evolution

FBF and Pumilio control expression of mRNAs that encode key regulators of development (Fig. 7). FBF restricts expression of *fem-3* mRNA so that sperm are only made in the most proximal region of the *C. elegans* germ line, as shown here, whereas Pumilio limits expression of *hunchback* mRNA to the anterior portion of the *Drosophila* embryo^{20,32,40,42}. In the absence of FBF, the pattern of sexual cell fates is lost from the hermaphrodite germ line and a uniform field of sperm is found in its place. Similarly, in the absence of Pumilio, the pattern of *hunchback* expression is lost and becomes uniform throughout the embryo³². Whereas FBF regulates *fem-3*, which directly controls the decision between spermatogenesis and oogenesis, Pumilio regulates *hunchback*, which then participates in the regulation of the more complex pattern of the fly embryo (reviewed in refs 6, 8–11).

The asymmetry imposed by Pumilio on *hunchback* expression is generated not by Pumilio alone, but in combination with another protein, Nanos, that is localized to the posterior pole of the embryo^{20,32,42,43}. By analogy, FBF may likewise collaborate with one or more partners to achieve the pattern of sexual fates in the *C. elegans* germ line. The best candidates for such FBF partners are the products of the *mog* genes, as they too are required for the sperm/oocyte switch^{33,34}.

XX *C. elegans* are self-fertilizing hermaphrodites, whereas XO animals are male. As a result, *C. elegans* can reproduce either by self-fertilization or by mating. Closely related nematode species are gonochoristic (male/female) species and reproduce only by mating (see ref. 44, for example). Because many branches of the nematode phylogenetic tree contain both hermaphroditic and gonochoristic species, and because hermaphroditic species have both males and hermaphrodites, evolution of a hermaphroditic sex from a female programme probably takes place with relative ease. This requires mechanisms to activate spermatogenesis and then to repress it. FBF

and *fem-3* play a central role in the requisite switch from male to female gametogenesis. An understanding of their regulation should illuminate this fundamental evolutionary change in reproductive strategy. □

Methods

Three-hybrid selection and screening. A derivative of yeast strain L40-coat¹⁹, containing pIIIA/*fem-3*-MS2, was transformed with the *C. elegans* library pRB1 (gift from R. Barstead). Transformants were plated onto synthetic media lacking leucine and histidine. 5 mM 3-aminotriazole was used to select for higher levels of activation of *HIS3*. About 140,000 transformants were screened. After a week, white colonies were picked and assayed for β -galactosidase activity, using either a filter assay with X-Gal as an indicator dye, or by liquid assays performed in triplicate using Galacton Plus as a luminescent substrate (Galacto-Light Plus assay kit from Tropix). One unit of enzyme activity (Fig. 2) is 10,000-fold greater than that defined by the supplier. Of 42 original *HIS3*⁺ transformants, six were white (*ADE2*⁺) and 36 were red or red-sectored (*ade2*⁻). Of the six white *HIS3*⁺ colonies, one was genuinely RNA-dependent; it was *fbf-1*. Subsequent screens of ~500,000 transformants yielded 18 RNA-dependent positives, of which three bound to *fem-3*- but not to IRE(iron-response element)-containing hybrid RNAs. Two of these positives were *fbf-2* (B. Kraemer and M.W., unpublished).

Plasmids encoding hybrid RNAs. Hybrid RNA vector: The vector encoding the hybrid RNA, designated pIII/MS2, also carries the *ADE2* and *URA3* genes. It was created by insertion of a *SpeI*–*SalI* fragment containing *S. cerevisiae ADE2* into the *NheI*–*XhoI* sites of pIII/MS2-2 (ref. 19). The hybrid RNAs were transcribed by RNA polymerase III (ref. 19).

Hybrid *fem-3* RNAs: To construct pIIIA/*fem-3*(+/MS2, pIIIA/*fem-3*(q96)/MS2, and pIIII/*fem-3*(ch8)/MS2, phosphorylated complementary oligonucleotides (sequence given in Fig. 1c) were annealed and ligated into pBluescript KS(+). A product with a tandem duplication of the *fem-3* region was inserted into the hybrid RNA vector. The two 37-nucleotide (nt) segments of the *fem-3* 3'UTR (Fig. 1c) are separated by a 4-nt spacer, UGGG. pIIII/*fem-3*(ch8)/MS2 does not contain the *ADE2* marker.

Hybrid RNA controls: IRE (iron response element)¹⁹, oligo(A)₃₀, and HIV-E (nt 224 to 798 of the HIV-1 isolate N14-3) were used as controls for sequence specificity. Each of these control hybrid RNAs interacts in the three-hybrid assay with its cognate protein (IRP, poly(A)-binding protein, and HIV Gag, respectively; D. SenGupta, B. Kraemer, S.F. and M.W., unpublished)¹⁹, in support of inability to interact with FBE.

Preparation of antibodies and immunofluorescence staining. To produce anti-FBF antibodies, rats were injected with purified glutathione-S-transferase (GST)-tagged FBF-1 protein (containing amino acids 121 to 614). Antibodies were purified over a matrix (Sterogene) carrying hexahistidine-tagged FBF-1 protein. To assay antibody specificity, L2 and L3 larvae were obtained from N2 hermaphrodites that were either uninjected or injected with *fbf-1* or *fbf-2* antisense RNAs. Larvae were washed and fixed as described⁴⁵. Siblings were checked for sterility as adults to ensure RNAi efficiency. To control for permeability, we used antibodies that recognize DNA (mAb030, Chemicon) and tubulin. To identify sperm, adult progeny from N2 hermaphrodites injected with *fbf-1* or F54C9.8 antisense RNAs were stained with SP56 antibodies, which recognize a sperm-specific protein in *C. elegans*²⁵, and with DAPI (4,6-diamidino-2-phenylindole), using the paraformaldehyde/dimethylformamide method (R. Lin, personal communication⁴⁵).

RNA-mediated interference (RNAi). Templates for *in vitro* transcription were cloned into pBluescript II KS(+) and transcribed using T7 RNA polymerase (Megascript T7 kit, Ambion). pBSII-fb1(−) carries *fbf-1* cDNA encoding amino acids 121–614 of FBF-1, pBSII-fb2(−) carries *fbf-2* cDNA encoding amino acids 1 to 630 of FBF-2, and pBSII-F54C9.8(−) includes amino acids 1 to 553 of F54C9.8. F54C9.8 DNA was obtained using PCR. RNA integrity was determined by gel electrophoresis; concentration was determined by UV spectrophotometry and confirmed by ethidium bromide staining. Injections used uncapped antisense RNA at a concentration of 5.0 mg ml^{−1} RNA, unless otherwise noted. The RNA used in Fig. 4 at 2.2 mg ml^{−1} was capped. Germ lines of progeny produced between 12 and 48 h after injection were examined for gamete differentiation either by Nomarski microscopy or fertility.

Preparation and analysis of *fbf-1* deletions. Deletions were prepared either

by digestion with restriction enzymes or by PCR. End-points were confirmed by restriction enzyme digests. Levels of *lacZ* expression were determined by filter assays, using three independent transformants for each DNA.

Isolation of full-length *fbf-1* and *fbf-2* cDNAs. Two oligo(dT)-primed cDNA libraries (AAE.1 and λ RB.1, gift from R. Barstead) were screened using standard procedures with a probe containing a fragment of *fbf-1* (nt 489–1,995). To obtain the 5' terminal portions of *fbf-1* and *fbf-2* cDNAs, we screened a random-primed *C. elegans* cDNA library, λ RB2 (gift from R. Barstead) with a fragment corresponding to nt 129–280 of *fbf-2*.

Genomic *fbf* sequences. *fbf-2* genomic sequence was determined in the *C. elegans* genome project²¹. The *fbf-2* gene is located on chromosome II near *lin-4*, and is present on the cosmid F21H12. To locate *fbf-1*, a polytene filter (poly2) corresponding to the *C. elegans* genome (gift from A. Coulson) was probed with *fbf-1* cDNA. The YACs identified by this method (Y56A9, Y45F1 and Y70E12L) correspond to a cluster of YACs on chromosome II that also map near *lin-4*. By performing PCR with primers specific to *fbf-1* and *fbf-2* using each YAC as template, we found that Y56A9 contains both *fbf-1* and *fbf-2*. Although both *fbf-1* and *fbf-2* are present on a single 450-kb YAC, they are not immediately adjacent: the genomic sequence of the *fbf-2* region reveals that it is flanked by unrelated genes (*C. elegans* Sequencing Consortium, personal communication).

Strains and genetics. Wild-type worms used were *C. elegans* variety Bristol, strain N2. Additional strains include: JK574 *fog2(q71)* V, JK255 *her-1(e1518)* *him-5 (e1490)* *dpy-21(e428ts)* V and CB3905 *unc-24(e138)* *fem-3 (e1996)* *+/+* *daf-15(m31)* *dpy-20(e128ts)* IV. These genes and mutations are described in ref. 46.

Defining the family of proteins containing Puf repeats and Csp consensus sequences. In the following consensus sequences, 'x' indicates any amino acid, and a slash ('/') separates amino acids of similar character. Certain Puf proteins are more closely related to one another, suggesting subfamilies: for example, FBF is more closely related to *C. elegans* F54C9.8 than to other Puf proteins.

Puf repeats and the Puf family: The NIH non-redundant database was scanned for matches to *fbf-1* and *pumilio* separately. Criteria for inclusion as a member of the family of Puf proteins included matches to a core consensus (D-x-F/Y-G/A-x-x-V/I-V/I/L-Q-K-x-V/I/L) within each repeat, and to the pattern of amino acids at positions 5 (NCNCNSN) and 6 (YRHRyXy) of the consensus in Puf2 through to Puf8 (Table 1). Proteins were assigned to the Puf protein family if they conformed to at least eight of the eleven criteria imposed by the Puf family core consensus above (eight amino-acid identities and three spacing). Repeats exhibiting only seven matches are indicated by a question mark in Fig. 6c. Several Puf repeats possess distinctive sequence features that distinguish them from the other repeats, but all share features of the overall core consensus³².

Received 1 August; accepted 17 October 1997

- Wickens, M., Kimble, J. & Strickland, S. In *Translational Control* (eds Hershey, J., Mathews, M. & Sonenberg, N.) 411–450 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1996).
- Sonenberg, N. mRNA translation: influence of the 5' and 3' untranslated regions. *Curr. Opin. Genet. Dev.* **4**, 310–315 (1994).
- Beelman, C. A. & Parker, R. Degradation of mRNA in eukaryotes. *Cell* **81**, 179–183 (1995).
- Wickens, M., Anderson, P. & Jackson, R. Life and death in the cytoplasm: messages from the 3' end. *Curr. Opin. Genet. Dev.* **7**, 220–232 (1997).
- Singer, R. RNA zipcodes for cytoplasmic addresses. *Curr. Biol.* **3**, 719–721 (1993).
- Curtis, D., Lehmann, R. & Zamore, P. D. Translational regulation and development. *Cell* **81**, 171–178 (1995).
- Anderson, P. & Kimble, J. In *C. elegans II* (eds Riddle, D. L., Blumenthal, T., Meyer, B. J. & Priess, J. R.) 185–208 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1997).
- St Johnston, D. & Nüsslein-Volhard, C. The origin of pattern and polarity in the *Drosophila* embryo. *Cell* **68**, 201–219 (1992).
- Macdonald, P. & Smibert, C. A. Translational regulation of maternal mRNAs. *Curr. Opin. Gen. Dev.* **6**, 403–407 (1996).
- Wharton, R. Regulated expression from maternal mRNAs in *Drosophila*. *Sem. Dev. Biol.* **3**, 391–397 (1992).
- St Johnston, D. The intracellular localization of messenger RNAs. *Cell* **81**, 161–170 (1995).
- Gebauer, R. & Richter, J. Synthesis and function of Mos: the control switch of vertebrate oocyte meiosis. *BioEssays* **19**, 23–28 (1997).

- Goodwin, E. B., Okkema, P. G., Evans, T. C. & Kimble, J. Translational regulation of *tra-2* by its 3' untranslated region controls sexual identity in *C. elegans*. *Cell* **75**, 329–339 (1993).
- Ahringer, J. & Kimble, J. Control of the sperm-oocyte switch in *Caenorhabditis elegans* hermaphrodites by the *fem-3* 3' untranslated region. *Nature* **349**, 346–348 (1991).
- Hodgkin, J. Sex determination in the nematode *C. elegans*: analysis of *tra-3* suppressors and characterization of *fem* genes. *Genetics* **114**, 15–52 (1986).
- Barton, M. K., Schedl, T. B. & Kimble, J. Gain-of-function mutations of *fem-3*, a sex-determination gene in *Caenorhabditis elegans*. *Genetics* **115**, 107–119 (1987).
- Ahringer, J., Rosenquist, T. A., Lawson, D. N. & Kimble, J. The *Caenorhabditis elegans* sex determining gene *fem-3* is regulated post-transcriptionally. *EMBO J.* **11**, 2303–2310 (1992).
- Ahringer, J. Post-transcriptional regulation of *fem-3*, a sex-determining gene of *C. elegans*. Ph.D. thesis, University of Wisconsin-Madison (1991).
- SenGupta, D. et al. A three-hybrid system to detect RNA-protein interactions *in vivo*. *Proc. Natl. Acad. Sci. USA* **93**, 8496–8501 (1996).
- Murata, Y. & Wharton, R. Binding of Pumilio to maternal hunchback mRNA is required for posterior patterning in *Drosophila* embryos. *Cell* **80**, 747–756 (1995).
- Waterston, R. H., Sulston, J. E. & Coulson, A. R. In *C. elegans II* (eds Riddle, D. L., Blumenthal, T., Meyer, B. J. & Priess, J. R.) 23–45 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1997).
- Austin, J. & Kimble, J. *glp-1* is required in the germ line for regulation of the decision between mitosis and meiosis in *C. elegans*. *Cell* **51**, 589–599 (1987).
- Guo, S. & Kemphues, K. J. *par-1*, a gene required for establishing polarity in *C. elegans* embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. *Cell* **81**, 611–620 (1995).
- Rocheleau, C. E. et al. Wnt signaling and an APC-related gene specify endoderm in early *C. elegans* embryos. *Cell* **90**, 707–716 (1997).
- Ward, S., Roberts, T., Strome, S., Pavalko, F. & Hogan, E. Monoclonal antibodies that recognize a polypeptide antigenic determinant shared by multiple *Caenorhabditis elegans* sperm-specific proteins. *J. Cell Biol.* **102**, 1778–1786 (1986).
- Meyer, B. J. In *C. elegans II* (eds Riddle, D. L., Blumenthal, T., Meyer, B. J. & Priess, J. R.) 209–240 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1997).
- Schedl, T. & Kimble, J. *fog-2*, a germ-line specific sex determination gene required for hermaphrodite spermatogenesis in *Caenorhabditis elegans*. *Genetics* **119**, 43–61 (1988).
- Hodgkin, J. More sex-determination mutants of *Caenorhabditis elegans*. *Genetics* **96**, 649–664 (1980).
- Fu, X.-d. The superfamily of arginine/serine-rich splicing factors. *RNA* **1**, 663–680 (1995).
- 't Harel, J. & Green, M. The SR protein family: pleiotropic functions in pre-mRNA splicing. *Trends Biochem. Sci.* **21**, 296–301 (1996).
- Macdonald, P. The *Drosophila pumilio* gene: an unusually long transcription unit and an unusual protein. *Development* **114**, 221–232 (1992).
- Barker, D., Wang, C., Moore, J., Dickinson, L. & Lehmann, R. *Pumilio* is essential for function but not for distribution of the *Drosophila* abdominal determinant Nanos. *Genes Dev.* **6**, 2312–2326 (1992).
- Graham, P. L. & Kimble, J. The *mog-1* gene is required for the switch from spermatogenesis to oogenesis in *Caenorhabditis elegans*. *Genetics* **133**, 919–931 (1993).
- Graham, P. L., Schedl, T. & Kimble, J. More *mog* genes that influence the switch from spermatogenesis to oogenesis in the hermaphrodite germ line of *Caenorhabditis elegans*. *Dev. Genet.* **14**, 471–484 (1993).
- Schedl, T., Graham, P. L., Barton, M. K. & Kimble, J. Analysis of the role of *tra-1* in germline sex determination in the nematode *Caenorhabditis elegans*. *Genetics* **123**, 755–769 (1989).
- Coglevina, M., Bertani, L., Klima, R., Zaccaria, P. & Bruschi, C. The DNA sequence of a 7941 bp fragment of the left arm of chromosome VII of *Saccharomyces cerevisiae* contains four open reading frames including the multicopy suppressor gene of the *pop2* mutation and a putative serine/threonine protein kinase gene. *Yeast* **11**, 767–774 (1995).
- Zamore, P., Williamson, J. & Lehmann, R. The *Pumilio* protein binds RNA through a conserved domain that defines a new class of RNA-binding protein. *RNA* (in the press).
- Birney, E., Kumar, S. & Krainer, A. Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Res.* **21**, 5803–5816 (1993).
- Nagai, K., Oubridge, C., Ito, N., Avis, J. & Evans, P. The RNP domain: a sequence-specific RNA-binding domain involved in processing and transport of RNA. *Trends Biochem. Sci.* **20**, 235–240 (1995).
- Wharton, R. P. & Struhl, G. RNA regulatory elements mediate control of *Drosophila* body pattern by the posterior morphogen *nanos*. *Cell* **67**, 955–967 (1991).
- Curtis, D. et al. A CCHC metal-binding domain in Nanos is essential for translational regulation. *EMBO J.* **16**, 834–843 (1997).
- Lehmann, R. & Nüsslein-Volhard, C. The maternal gene *nanos* has a central role in posterior pattern formation in the *Drosophila* embryo. *Development* **112**, 679–691 (1991).
- Wang, C., Dickinson, L. & Lehmann, R. Genetics of *nanos* localization in *Drosophila*. *Dev. Dynam.* **199**, 103–115 (1994).
- Fitch, D. H. A. & Thomas, W. K. In *C. elegans II* (eds Riddle, D. L., Blumenthal, T., Meyer, B. J. & Priess, J. R.) 815–850 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1997).
- Crittenden, S.-L., Troemel, E. R., Evans, T. C. & Kimble, J. *GLP-1* is localized to the mitotic region of the *C. elegans* germ line. *Development* **120**, 2901–2911 (1994).
- Hodgkin, J. In *C. elegans II* (eds Riddle, D. L., Blumenthal, T., Meyer, B. J. & Priess, J. R.) 881–1047 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1997).

Acknowledgements. We thank R. Lehmann, C. Mello, J. Priess, R. Wharton and J. Williamson for communicating results before publication and for discussion; P. Anderson for comments on the manuscript; R. Barstead for *C. elegans* cDNA libraries; A. Coulson for YAC grids; the *C. elegans* sequencing consortium for unpublished genomic sequences; S. Ward for anti-sperm antibodies; and U.W. Biochemistry Media Laboratory for help with preparing the figures. DNA sequencing was performed at the UW Biotechnology Center. This work was supported by grants from the NIH to J.K. and M.W., from the NSF Science and Technology Center (Seattle) to S.F. and M.W., and by EMBO and Fonds National Suisse fellowships to A.P., J.K. and S.F. are investigators with the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to M.W. (e-mail: wickens@biochem.wisc.edu).

The Evershed effect in sunspots as a siphon flow along a magnetic flux tube

Benjamín Montesinos* & John H. Thomas†

* LAEFF/INTA, PO Box 50727, E-28080 Madrid, Spain, and IAA/CSIC, PO Box 3004, E-18080 Granada, Spain

† Department of Mechanical Engineering, Department of Physics & Astronomy, and C.E.K. Mees Observatory, University of Rochester, Rochester, New York 14627, USA

The Evershed effect¹—a wavelength shift and profile asymmetry in the spectral lines observed from the outer regions of sunspots (the penumbra)—has been interpreted as a radial outflow of gas from the sunspot, but the dynamics of the flow have not been fully understood². Although the Evershed effect seems to stop abruptly at the outer edge of the penumbra, the outflow itself must continue, though tracing its path has proved difficult. Theoretical^{3,4} and observational^{5–7} studies have suggested that much of the continuing flow may follow magnetic field lines that go below the visible surface of the Sun at or just beyond the edge of the penumbra, and recent observations have now confirmed this picture⁸. Here we show, using theoretical calculations based on a more realistic model, that the flow acts like a siphon^{3,9–12} which is driven along a magnetic flux tube by the pressure drop between the endpoints of the tube.

Over the past decade, through theoretical advances¹³ and high-resolution observations, we have learned that a sunspot penumbra is a deep structure containing a significant fraction of the sunspot's emerging magnetic flux. The penumbral magnetic field has a fluted structure, with radial spokes in which the magnetic field is inclined to the horizontal (at $\sim 40^\circ$) alternating with regions in which the magnetic field is very nearly horizontal^{14–16}. The photospheric Evershed flow is largely confined to thin, elevated radial channels¹⁷ lying in regions where the magnetic field is nearly horizontal.

The configuration of the magnetic field and Evershed flow beyond the outer penumbral boundary has been less well understood. Because of the relatively high electrical conductivity of the gas, the Evershed flow is constrained to be along the direction of the magnetic field lines. We know that much of the magnetic flux in the penumbra continues radially outward beyond the penumbral boundary in the form of a magnetic canopy elevated above the surrounding quiet photosphere^{18,19}. Indeed, a continuation of the outward Evershed flow along this magnetic canopy has been detected²⁰, but the mass flux associated with this canopy flow is only a fraction of the total associated with the Evershed effect²⁰. Another possibility, suggested by the siphon-flow model, is that many of the arched magnetic flux tubes carrying the Evershed flow return to the solar surface and dive below it at points either just within or just outside the outer boundary of the penumbra^{3,4}. Recent observations have provided increasing evidence that the Evershed flow can be traced along individual magnetic channels which end in intense magnetic features of opposite polarity outside the sunspot^{5–7}, or which dive below the visible surface within the penumbra itself²¹. The very recent observations of Westendorp Plaza *et al.*⁸ have now shown convincingly that both of these configurations do indeed occur. Their observations are based on a careful inversion of the full Stokes profiles of magnetically sensitive spectral lines (Fe I 620.15 and 630.25 nm) which traces the magnetic field and flow velocity as functions of height, and which clearly connects the Evershed flow to downflows in patches of opposite magnetic polarity near the outer edge of the sunspot.

This configuration for the Evershed flow is precisely what is

predicted by the siphon-flow mechanism, in which a flow along an arched magnetic flux tube is driven by a pressure drop between the two footpoints of the arch. We have already presented siphon-flow calculations for arched flux tubes lying entirely within the penumbra³. Here we present new calculations based on a more realistic penumbral model which also allows us to follow flux tubes that cross the penumbra–photosphere boundary and return to the surface in the photosphere outside the sunspot. Thus, we are now able to model both types of flow-carrying flux tubes observed by Westendorp Plaza *et al.*⁸.

Our new model of the ambient atmosphere consists of a magnetic penumbral atmosphere in mechanical equilibrium with an external field-free atmosphere representing the surrounding quiet photosphere. The penumbral magnetic field is assumed to be force-free and we neglect horizontal variations in the penumbra. The penumbral and photospheric atmospheres are each vertically stratified, and the total pressure (gas pressure plus magnetic pressure) in the penumbra equals the gas pressure in the photosphere everywhere

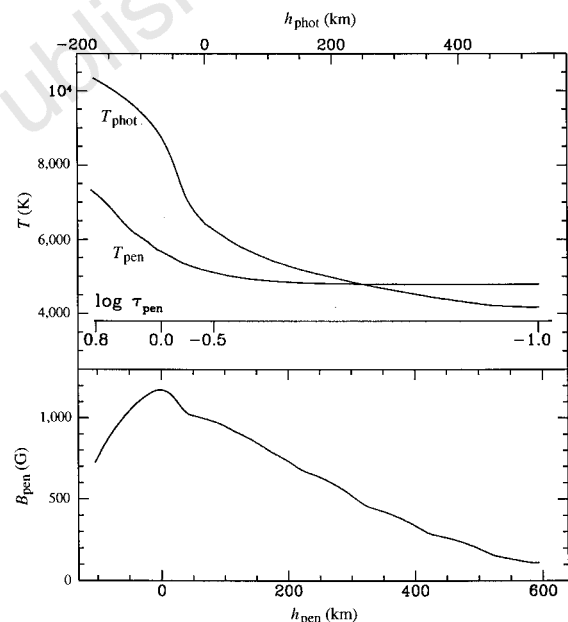


Figure 1 The distribution with height of the penumbral magnetic field strength B_{pen} , penumbral temperature T_{pen} , and photospheric temperature T_{phot} in the model ambient atmosphere. The geometric heights h_{pen} and h_{phot} are offset by a Wilson depression of 70 km. Also shown is the scale of the optical depth τ_{pen} in the penumbra. Our model of the quiet photosphere consists of the VAL C model²⁸ of the visible layers matched to Spruit's model²⁹ of the underlying convection zone. The penumbral atmosphere has an ambient magnetic field of strength $B_{\text{pen}}(h)$, which varies with geometric height h and is in pressure equilibrium with the quiet photospheric atmosphere at the interface: $p_{\text{pen}} + B^2/8\pi = p_{\text{phot}}$. The penumbral temperature at optical depth τ is related to the quiet photospheric temperature at the same optical depth by the empirical relation $\Delta\theta = 5,040[1/T_{\text{pen}}(\tau) - 1/T_{\text{phot}}(\tau)]$ of Kjeldseth Moe and Maltby³⁰, with the value $\Delta\theta = 0.093$ appropriate for the dark penumbral filaments. We apply this relation only for optical depths less than $\log\tau_{\text{pen}} = 0.8$, which corresponds to geometric height $h_{\text{pen}} = -100$ km in our model. (We use this height in the penumbra as the starting point for our siphon-flow calculations, where we specify the flow speed, tube diameter, and inclination.) To obtain the temperature and density as functions of h in the ambient penumbra, we integrate the hydrostatic equation $dp = -\rho g dh$ and the relation $d\tau = -\rho\kappa dh$, where ρ is the mass density and κ is the grey Rosseland opacity, after first setting the value of the penumbral magnetic field strength at $\tau = 1.0$ to be 1,200 G and the offset between geometric heights in the photosphere and penumbra (that is, the Wilson depression of the penumbra) to be $h_{\text{phot}} - h_{\text{pen}} = 70$ km. Our choice of tabled values of the opacity κ as a function of temperature and density is described elsewhere²².

along the interface (magnetopause) between the two. Figure 1 shows the distribution with height of the temperature in the ambient penumbra, the temperature in the ambient photosphere, and the magnetic field strength in the ambient penumbra; further details of our model ambient atmosphere are given in the legend of Fig. 1.

Our formulation of the steady siphon flow of a zero-viscosity gas along a thin, arched magnetic flux tube, and our method of numerical integration of the basic equations, have been presented elsewhere^{3,22}. We calculate simultaneously the steady flow speed along the tube, the cross-sectional area (and hence magnetic field strength) of the tube, and the equilibrium path of the flux tube through the surrounding external atmosphere. The calculations include the effects of radiative transfer between the flux tube and its surroundings, and variable ionization of the flowing gas. Our treatment of the radiative transfer is valid both in regions where the flux tube is optically thin and in regions where it is optically thick.

Figure 2 shows four representative examples of computed siphon flows along flux tubes emerging within the penumbra: one lying entirely within the penumbra (case a), and the others (cases b–d) extending across the interface into the surrounding quiet photosphere. The bottom panel shows the equilibrium path of the flux tube axis through the atmosphere (height h versus horizontal distance x). The equilibrium path of the flux tube is determined by a balance among the magnetic buoyancy force, the net magnetic tension force, and the centrifugal force of the flow along the curved path. We note that in cases b–d, where the flux tube extends outside the penumbra into the surrounding photosphere, the curvature of the arch is greater outside the penumbra in order for the net magnetic tension force to balance the greater magnetic buoyancy of the flux tube in the field-free surroundings. The configurations of the arched flux tubes in Fig. 2 strongly resemble those sketched by Westendorp Plaza *et al.*⁸ (their Fig. 2) on the basis of their observations.

The middle and upper panels of Fig. 2 show the flow speed and magnetic field strength along the flux tube. For all four cases shown

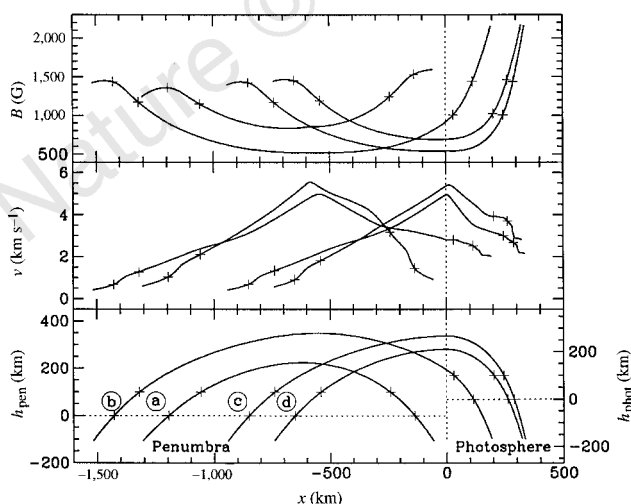


Figure 2 Four examples of computed siphon flows. We show the equilibrium path (height h versus horizontal distance x) of the axis of the thin magnetic flux tube through the ambient atmosphere (bottom panel), and the flow speed (middle panel) and magnetic field strength (top panel) along the flux tube. The horizontal dotted lines represent the level of continuum optical depth unity in the penumbra and in the surrounding photosphere; the vertical dotted line indicates the outer boundary of the penumbra. The crosses mark the upstream and downstream points where the geometric height (h_{pen} or h_{phot} , as appropriate) equals 0 and 100 km. Case a is for an arched flux tube lying entirely within the penumbra, whereas cases b–d are for flux tubes which return to the solar surface outside the sunspot. Both such cases have been observed by Westendorp Plaza *et al.*⁸.

here, the flow speeds at the downstream footpoint (at continuum optical depth unity) are about $3\text{--}4\text{ km s}^{-1}$, in good agreement with the speeds measured by Westendorp *et al.*⁸. For the flux tubes extending out into the surrounding photosphere, the magnetic field strength at the downstream footpoint in the photosphere (at optical depth unity) is of the order of 1,400 G, and at a height of 100 km (where the field strength is typically measured in the core of a line) it is of the order of 1,200 G. These values are typical of the field strengths measured in intense magnetic elements in the photosphere.

In a siphon flow, if we compare footpoints at the same absolute height h , the magnetic field strength is always higher at the downstream footpoint, where the internal gas pressure is necessarily lower. But because of the Wilson depression of the penumbra (due to its greater transparency than the photosphere outside the sunspot), if we make this comparison at the same relative geometric height in the penumbra (h_{pen}) and photosphere (h_{phot}), the magnetic field strength may be the same or lower than that at the penumbral footpoint. For example, for case c (Fig. 2) the field strength is 1,424 G at the upstream footpoint (at $h_{\text{pen}} = 0$) and 1,439 G at the downstream footpoint (at $h_{\text{phot}} = 0$); 100 km above that, the corresponding values are 1,169 G at the upstream point (at

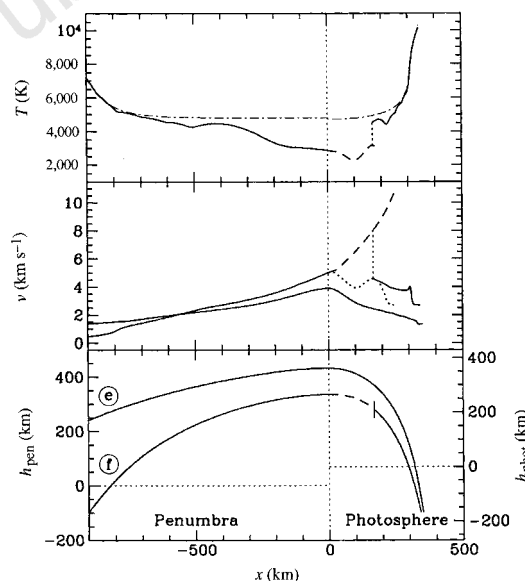


Figure 3 Examples of a slower subcritical flow (case e) and a critical flow with a standing tube shock in the downstream leg of the arch (case f). As in Fig. 2, the bottom panel shows the equilibrium path of the flux-tube axis, and the middle panel shows the distribution of flow speed along the tube. For the critical flow, supercritical speeds are indicated by a dashed curve. The critical flow undergoes a smooth transition from subcritical to supercritical speed at the top of the arch, and then an abrupt transition from supercritical to subcritical speed at the standing tube shock (indicated by the vertical dotted line segment) in the downstream leg of the arch. The position of the tube shock in a critical flow depends on the 'back-pressure' at the downstream footpoint. Here we have shown just one example; for lower (higher) backpressure, the shock will move to a position further downstream (upstream). The locus of the flow speed on the downstream side of the tube shock for different back-pressure is shown as a dotted curve in the middle panel. The top panel shows the distribution of internal temperature (solid, dashed, and dotted curves) and external temperature (dot-dash curve) along the flux tube for the critical flow. We note that the gas in the flux tube is cooler than the surroundings along the higher portions of the arch, where the radiative exchange is less efficient and the flow is nearly adiabatic. As a result of the lower temperature and density inside the flux tube, the upper portions of the flux tube arch are more transparent than the surroundings. At lower elevations, efficient radiative coupling keeps gas in the flux tube at nearly the same temperature as the surroundings.

$h_{\text{pen}} = 100 \text{ km}$) and $1,007 \text{ G}$ at the downstream footpoint (at $h_{\text{phot}} = 100 \text{ km}$). As the magnetic response height for measurements of the magnetic field strength in a photospheric spectral line (such as $\text{Fe I } 630.15 \text{ or } 630.25 \text{ nm}$) lies roughly in the range $0\text{--}100 \text{ km}$, these values show that observations can give the same or smaller magnetic field strength at the downstream "footpoint" of the flux-tube arch, even though the field strength is actually higher at the true downstream footpoint located at the same absolute height (or more precisely, at the same gravitational potential) as the upstream footpoint.

The flux tubes shown in Fig. 2 rise to a height of some $200\text{--}350 \text{ km}$ above the visible surface in the penumbra, and their diameters are typically $50\text{--}70 \text{ km}$ at the top of the arch. This configuration of a thin (diameter $\sim 0.1 \text{ arcsec}$), elevated flow channel in the penumbra is just what is seen in high-resolution observations^{7,17}. Changing the tube diameter affects the computed flow only weakly, by changing the radiative coupling with the surroundings. The horizontal span of the arches in Fig. 2 is comparable to, but perhaps slightly shorter than, that which can be inferred from observations^{5–8}. Longer arches can also be produced by our model, but these generally have somewhat lower flow speeds.

The siphon flows shown in Fig. 2 are each slightly subcritical. For a smaller pressure drop between the footpoints, the flow will be subcritical with lower flow speeds throughout. For a larger pressure drop, there will be a critical flow which undergoes a smooth transition to supercritical speed (that is, speed greater than the local 'tube speed' c_t , the speed of propagation of a pressure pulse in the tube) before decelerating suddenly to subcritical speed at a standing tube shock in the downstream leg of the arch²³. Figure 3 shows one example of each of these kinds of siphon flows. Case e, a lower-speed subcritical flow, has a smoother, less peaked distribution of flow speed along the tube, compared to the slightly subcritical flows shown in Fig. 2. Case f, a critical flow, passes through a critical point at the top of the arch and attains flow speeds of up to 8 km s^{-1} before decelerating to subcritical speed at the standing tube shock. Higher speeds of this order are regularly seen in high-resolution observations of the Evershed flow.

Where do the magnetic field lines of the Evershed flux tubes go after submerging below the surface near the outer penumbral boundary? In our steady siphon-flow model, the tubes must be anchored below the surface in some way in order to provide a radial inward force to sustain the equilibrium. If this anchoring force is relaxed, then the downstream footpoint will move radially outward. This suggests that the downstream footpoints of the Evershed flux tubes may correspond to the 'moving magnetic features' that drift radially outward across the 'moat' surrounding the sunspot, and the downflows sometimes seen in these features²⁴ may just be the continuation of the Evershed flow.

The siphon-flow mechanism can also account for the Evershed flow that is seen to continue radially outward along the base of the magnetic canopy outside the sunspot¹², and it can account naturally for the reversed chromospheric Evershed flow along higher arched flux tubes emanating from the sunspot umbra²⁵. Indeed, it now seems that the siphon-flow model gives a quite satisfactory theoretical explanation of most of the salient observed features of the Evershed flow. More remains to be done, however. We still do not know what determines the spatial distribution of the Evershed flux tubes in the penumbra. Also, it is now known that the Evershed flow is not strictly steady (as in our model); the flows along individual channels are transient with a timescale of the order of $10\text{--}15 \text{ minutes}$ (refs 21, 26). Siphon flows on the Sun will also be transient in nature, because of changing conditions at the footpoints¹² or because of convective interchange of flux tubes in the penumbra²⁷. But these changes occur on a convective timescale of 10 minutes or more, longer than the timescale for the flow to adjust to these changes, so we can anticipate that time-dependent siphon flows will be quasi-steady, and very much like the steady flows presented here. □

Received 13 August; accepted 22 October 1997.

1. Evershed, J. Radial movement in Sun-spots. *Mon. Not. R. Astron. Soc.* **69**, 454–457 (1909).
2. Thomas, J. H. & Weiss, N. O. in *Sunspots: Theory and Observations* (eds Thomas, J. H. & Weiss, N. O.) 3–59 (Kluwer, Dordrecht, 1992).
3. Thomas, J. H. & Montesinos, B. A siphon-flow model of the photospheric Evershed effect in a sunspot. *Astrophys. J.* **407**, 398–401 (1993).
4. Thomas, J. H. in *Solar and Astrophysical Magnetohydrodynamic Flows* (ed. Tsinganos, K. C.) 39–60 (Kluwer, Dordrecht, 1996).
5. Börner, P. & Kneer, F. High resolution observations of the Evershed flow. *Astron. Astrophys.* **259**, 307–312 (1992).
6. Rimmele, T. R. Sun center observations of the Evershed effect. *Astrophys. J.* **445**, 511–516 (1995).
7. Stanchfield, D. C. H., Thomas, J. H. & Lites, B. W. The vector magnetic field, Evershed flow, and intensity in a sunspot. *Astrophys. J.* **477**, 485–494 (1997).
8. Westendorp Plaza, C. *et al.* Evidence for an Evershed flow returning into the Sun. *Nature* **389**, 47–49 (1997).
9. Meyer, F. & Schmidt, H. U. Magnetisch ausgerichtete Strömungen zwischen Sonnenflecken. *Z. Angew. Math. Mech.* **48**, 218–221 (1968).
10. Thomas, J. H. Siphon flows in isolated magnetic flux tubes. *Astrophys. J.* **333**, 407–419 (1988).
11. Degenhardt, D. Stationary siphon flows in thin magnetic flux tubes. II. Radiative heat exchange with the surroundings. *Astron. Astrophys.* **248**, 637–646 (1991).
12. Thomas, J. H. in *Solar Surface Magnetism* (eds Rutten, R. J. & Schrijver, C. J.) 219–235 (Kluwer, Dordrecht, 1994).
13. Schmidt, H. U. Sunspots. *Geophys. Astrophys. Fluid Dyn.* **62**, 249–270 (1991).
14. Degenhardt, D. & Wiehr, E. Spatial variation of the magnetic field inclination in a sunspot penumbra. *Astron. Astrophys.* **252**, 821–826 (1991).
15. Title, A. M. *et al.* On the magnetic and velocity field geometry of simple sunspots. *Astrophys. J.* **403**, 780–796 (1993).
16. Lites, B. W., Elmore, D. F., Seagraves, P. & Skumanich, A. P. Stokes profile analysis and vector magnetic fields. VI. Fine scale structure of a sunspot. *Astrophys. J.* **418**, 928–942 (1993).
17. Rimmele, T. R. Evidence for thin elevated Evershed channels. *Astron. Astrophys.* **298**, 260–276 (1995).
18. Giovanelli, R. & Jones, H. P. Three-dimensional structure of atmospheric magnetic fields. *Solar Phys.* **79**, 267–278 (1982).
19. Solanki, S. K., Rüedi, I. & Livingston, W. Infrared lines as probes of solar magnetic features. V. The magnetic structure of a simple sunspot and its canopy. *Astron. Astrophys.* **263**, 399–350 (1992).
20. Solanki, S. K., Montavon, C. A. P. & Livingston, W. Infrared lines as probes of solar magnetic features. VII. On the nature of the Evershed effect in sunspots. *Astron. Astrophys.* **283**, 221–231 (1994).
21. Shine, R. A. *et al.* High-resolution observations of the Evershed effect in sunspots. *Astrophys. J.* **430**, 413–424 (1994).
22. Montesinos, B. & Thomas, J. H. Siphon flows in isolated magnetic flux tubes. V. Radiative flows with variable ionization. *Astrophys. J.* **402**, 314–325 (1993).
23. Thomas, J. H. & Montesinos, B. Siphon flows in isolated magnetic flux tubes. IV. Critical flows with standing tube shocks. *Astrophys. J.* **375**, 404–413 (1991).
24. Nye, A. H., Thomas, J. H. & Cram, L. E. Dynamical phenomena in sunspots. II. A moving magnetic feature. *Astrophys. J.* **285**, 381–385 (1984).
25. Spruit, H. C. in *The Physics of Sunspots* (eds Cram, L. E. & Thomas, J. H.) 359–368 (Sacramento Peak Observatory, 1991).
26. Rimmele, T. R. On the temporal behaviour of the Evershed effect. *Astron. Astrophys.* **290**, 972–982 (1994).
27. Schlichenmaier, R., Jahn, K. & Schmidt, H. U. in *Advances in the Physics of Sunspots* (eds Schmieder, B., del Toro Iniesta, J. C. & Vázquez, M.) 140–144 (ASP Conf. Ser. Vol. 118, Astron. Soc. Pacific, San Francisco, 1997).
28. Vernazza, J. E., Avrett, E. H. & Loeser, R. Structure of the solar chromosphere. III. Models of the EUV brightness components of the quiet Sun. *Astrophys. J. Suppl. Ser.* **45**, 635–725 (1981).
29. Spruit, H. C. A model of the solar convection zone. *Solar Phys.* **34**, 277–290 (1974).
30. Kjeldseth Moe, O. & Maltby, P. The temperature of penumbral filaments. *Solar Phys.* **36**, 101–108 (1974).

Acknowledgements. We thank the authors of ref. 8 for discussions of their observations. B.M. was supported by a grant from the Spanish DGICYT, and J.H.T. was supported by a grant from NASA.

Correspondence should be addressed to J.H.T. (email: thomas@astro.me.rochester.edu).

Critical currents approaching the depairing limit at a twin boundary in $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$

Ivan Maggio-Aprile, Christophe Renner, Andreas Erb, Eric Walker & Øystein Fischer

Département de Physique de la Matière Condensée, Université de Genève, 24, Quai E.-Ansermet, CH-1211 Genève 4, Switzerland

Interest in vortex matter has risen considerably since the discovery of the high-temperature superconductors, which exhibit magnetic vortex states that are especially rich and complex¹. The global behaviour of magnetic vortices in nearly perfect crystals—such as melting of the vortex lattice^{2–5}—has been much studied, but of more technological relevance is the influence on the vortex states of the various structural defects present in most practical superconductors. An important example of such a defect is the twin boundary present in twinned orthorhombic crystals of

YBa₂Cu₃O_{7-δ} (YBCO). Studies of such samples using magnetic-field-sensitive probes^{6–9} have suggested that the twin boundary plays an important role in pinning the vortices and so enhancing the currents that YBCO can support while remaining superconducting. But the low spatial resolution of these techniques does not permit these effects to be studied at the scale of the vortices or boundaries themselves. Scanning tunnelling spectroscopy offers a means of circumventing these problems of resolution^{10–12}, as it directly probes the superconducting order parameter at nanometre length scales. Here we use this technique to investigate the importance of twin boundaries in YBCO. In particular, we observe an unexpectedly large pinning strength for perpendicular vortex flux across the boundary, which implies that the critical current that can be supported along the boundary approaches the theoretical 'depairing' limit.

The tunnelling experiments presented here were performed on an ultra pure twinned single crystal of YBCO¹³, optimally oxygenated ($T_c \approx 91$ K, transition width ~ 0.2 K), without any special surface conditioning. Our experimental set-up is described in detail elsewhere^{14,15}. Spectroscopic measurements were made by sweeping the tunnel voltage, maintaining a constant distance from tip to sample, and extracting the tunnelling conductance using a lock-in amplifier (1.0 to 2.0 mV modulation). The tunnelling resistance is maintained at high values (~ 1 GΩ) to ensure a sufficiently large tip-sample distance. All measurements were made at $T = 4.2$ K, in the presence of a He exchange gas pressure of 2×10^{-3} mbar. The STM tips were made of chemically etched pure iridium wire.

The magnetic field is applied perpendicular to the YBCO basal *ab*-plane. Vortex lattice imaging relies on local variations of the quasiparticle density of states. When the tip is positioned outside the vortex cores, the conductance spectra are basically identical to those measured in the Meissner state¹². On the other hand, we measure completely different tunnelling spectra when the tip enters

the core region. An image of the flux line lattice is obtained by mapping these systematic changes in real space.

The vortex lattice for a crystal cooled in a field of 3 T is shown in the $170 \text{ nm} \times 170 \text{ nm}^2$ image of Fig. 1a, where the *x*- and *y*-axes coincide with the *a* and *b* directions of the crystal. In each domain (delimited by the dark line), the vortex lattice presents a tendency towards a short-range order with oblique symmetry similar to the vortex structure observed previously¹². The anisotropic shape of the cores is visible in both domains. In the area to the left, this elongation is along the *x*-axis of the image, whereas to the right of the dark line it is along the *y*-axis. A Fourier analysis of the image reveals that the oblique lattice symmetry also rotates by 90° between the domains. We conclude that the domains have inverted *a* and *b* crystallographic axes, and that the dark line, aligned with either the (110) or ($\bar{1}\bar{1}0$) crystallographic axes, is a signature of the twin boundary. To our knowledge, this is the first observation of a twin boundary using a scanning tunnelling microscope. But why does the twin boundary appear this way in the spectroscopic image? The most straightforward explanation is that a large number of flux lines are trapped along the twin boundary. This is consistent with the observation that on both sides of the boundary vortex cores are repelled to a finite distance from the line.

In Fig. 1a, the flux lines close to the boundary show a tendency to align themselves parallel to it. This tendency disappears at some distance from the boundary, where the rows align along other directions. Thus, the vortex lattice seems to be perturbed in the immediate vicinity of the twin boundary, but relaxes back to its unperturbed shape at a distance of a few vortex lattice spacings from it. This is apparently in agreement with numerical simulations¹⁶, but not with neutron diffraction measurements which suggest a long-range influence of the boundary on the vortex lattice orientation¹⁷.

Reducing the field from 3 to 1.5 T produces a strikingly different image. In a $150 \times 150 \text{ nm}^2$ image, taken at roughly the same

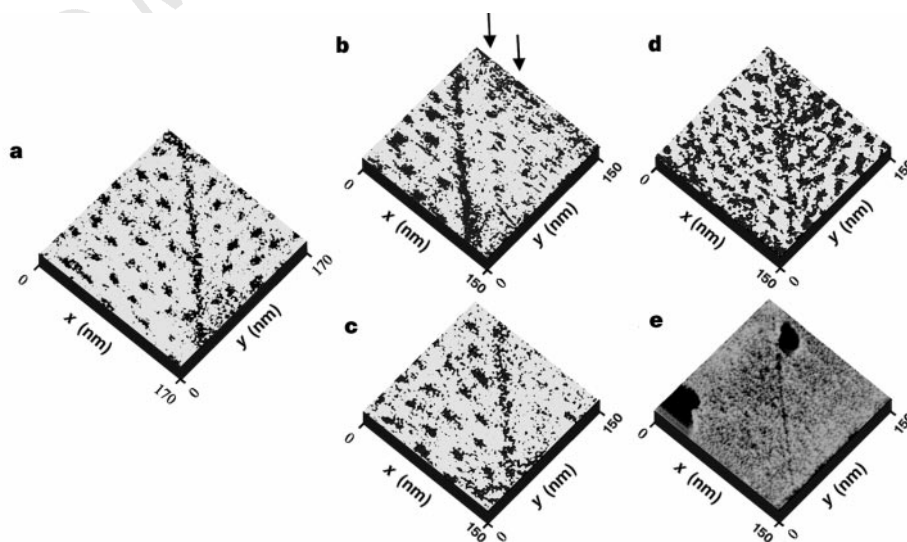


Figure 1 **a** to **d**, Spectroscopic images of the vortex lattice on the (001) YBCO single-crystal surface at $T = 4.2$ K, with the magnetic field applied parallel to the *c*-axis. The grey scale is defined by the ratio of the conductance measured at 20 meV to the conductance measured at zero bias. Grey tones varying from clear to dark correspond to ratios ranging from high to low. Note that for experimental reasons, the centre of the different images presented here are not exactly at the same position; this puts the twin boundary (TB) at slightly different positions relative to the centre in the various images. **a**, $170 \times 170 \text{ nm}^2$ image taken at 3 T (field-cooled). Both domains are filled with a nearly equal density of flux lines, and the 90° rotation of the *ab*-plane anisotropy is observed across the TB. **b**, $150 \times 150 \text{ nm}^2$ image taken 12 hours after the field was reduced from 3 to

1.5 T. The arrows indicate the vortex movements observed in the domain to the right, forming non-continuous lines extending in a direction parallel to the TB. **c**, Three days after field reduction, no more flux lines can be detected throughout the domain to the right over at least 80 nm. **d**, $150 \times 150 \text{ nm}^2$ image taken after a 3 T–1.5 T–0 T–6 T field cycle. Both domains show a high density of flux lines, and a flux gradient is measured across the TB. **e**, Topographic image of the YBCO surface taken simultaneously with **d**. The TB appears as a narrow structure about 0.1 nm deep. Note that the width of this line is much smaller than the width of the dark line in the spectroscopic images, giving further support to the interpretation that the dark line comes from a high density of vortices along the TB.

location 12 hours after the field was reduced, the boundary is still visible (Fig. 1b). However, although the domain to the left shows a high density of vortices, the domain to the right no longer shows any clear presence of cores, only low-contrast variations forming non-continuous lines extending parallel to the boundary (indicated with arrows in Fig. 1b). This is precisely what one expects to see if the vortices move from one site to another during the scan and if the vortex sites are only partly occupied. Whereas these movements are extremely slow in the domain to the left, the core displacements are much faster to the right. Whether the vortices trapped in the twin boundary move along it is an open question. Earlier magneto-optical investigations at low fields suggest that this might be the case if the Lorentz force is mainly parallel to the boundary^{7,8}. From our study we can neither confirm nor exclude that some movement of this type takes place. However, in the area we have investigated, this effect cannot be an effective channel for flux entry or exit, as we observe extreme field gradients across the TB.

Three days after field reduction, no signature of a vortex lattice can be detected through the domain on the right-hand side (Fig. 1c). We therefore conclude that there is a vortex-free region in the domain extending to the right over a distance of at least 80 nm. The average flux line density extracted from the domain to the left corresponds to a local field of about 2.5 T, noticeably higher than the applied field. These observations clearly suggest that the boundary acts as a very efficient barrier against flux-line movement in a direction orthogonal to the boundary plane, resulting in a pile-up of vortices on one side, and a shadowing effect on the other. As a consequence, very large currents are expected to flow close to the twin boundary.

Reducing the field to zero produces an image that is essentially the

same as the one seen in Fig. 1c. When the field is increased to 6 T, we recover a situation where both domains show a roughly equal density of flux lines (corresponding to a local field of about 6 T), as shown in the $150 \times 150 \text{ nm}^2$ image in Fig. 1d. The boundary is still clearly visible, as are individual flux lines. We note that the distance separating the twin boundary from the first row of vortices to the left ($\sim 20 \text{ nm}$) is larger than the distance to the right ($\sim 14 \text{ nm}$). The corresponding flux gradient is thus reversed when compared with the one observed in Fig. 1c with 1.5 T. Again here, we suggest that the boundary is at the origin of this gradient, behaving as a strong barrier when the vortices try to cross it.

The boundary seen in spectroscopy in Fig. 1a–d also shows up in topographic images as a trench about 0.1 nm deep and 2–3 nm broad. Such an image, recorded simultaneously with that of Fig. 1d, is illustrated in Fig. 1e. The rest of the image shows an atomically flat plateau with two small areas one unit cell below the plateau. It is possible that the narrow trench reflects a lower density of states along the boundary rather than a 'real' topographic effect.

Our observations show that the twin boundary acts as a strong pinning centre, and consequently a large number of vortices are trapped along it. An estimate for the pinning strength can be obtained from the observation, in the right-hand side of Fig. 1c of a vortex-free region of at least 80 nm ($\sim \lambda/2$ where λ is $\sim 150 \text{ nm}$ for YBCO). However, on the left, the average field is more than twice the thermodynamic critical field ($H_c = 1.1 \text{ T}$). Thus the screening current along the twin boundary must be of the order of the Ginzburg–Landau depairing current $j_0 = 4H_c(T)/(3\sqrt{6}\pi\lambda(T)) \approx 3 \times 10^8 \text{ A cm}^{-2}$.

When aiming to quantify and understand how the gradients are established, we find that the essential physics can be obtained from a simple model which ignores the anisotropy and the exact shape of the vortex lattice. We consider two domains separated by a twin boundary along which is confined a high density of trapped vortices to be determined by numerical simulation. Inside the two domains, there are no pinning centres. As we are only interested in field gradients perpendicular to the boundary, we assume that the vortices are arranged in rows parallel to the boundary. The fields far to the left and right are set to a specific value. We then assume that the distance between vortices along a row is equal to the distance between the rows far from the boundary, and we determine the position of each vortex row by the condition that the Lorentz force perpendicular to the row, produced by all the other rows, is zero. As we are looking at an extreme type II superconductor and $H_{c1} \ll H \ll H_{c2}$, this calculation can be carried out simply by using the London equation for a vortex lattice:

$$\lambda^2 \nabla_{xy}^2 B_z(x, y) - B_z(x, y) = \sum_i \phi_0 \delta(x - x_i, y - y_i)$$

where the summation runs over all vortex sites, ϕ_0 is the flux quantum $= h/2e = 2.07 \times 10^{-15} \text{ T m}^2$ and we assume $\lambda = 150 \text{ nm}$. The problem reduces to a one-dimensional case obtained by integrating the London equation parallel to the twin boundary. The positions of the individual vortex rows are calculated by an iterative process, for different conditions.

In the situation shown in Fig. 1a, where the fields in both domains are equal, our calculation gives a symmetric distribution of vortices across the boundary. When the field is lowered on the right-hand side, the vortices on this side get pushed away by the higher density of vortices on the left-hand side. At the same time the vortices on the left get pushed towards the boundary, unless a sufficiently high number of flux lines are present in the boundary. To simulate the situation shown in Fig. 1c we determine the field difference between the two domains from the assumption that the first row of vortices to the right is situated at least 80 nm from the boundary, which is the minimum distance consistent with our measurements. This occurs

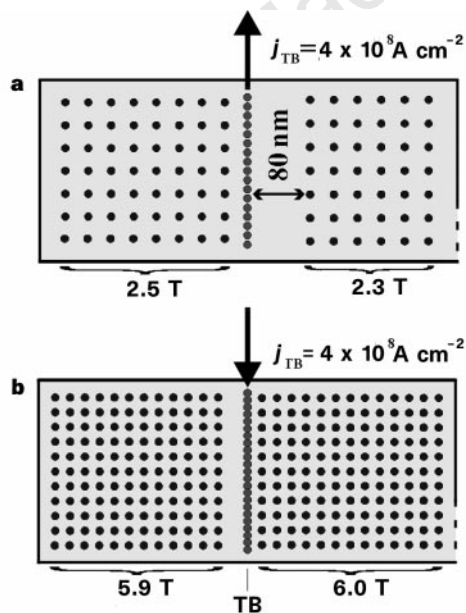


Figure 2 Simulation of the behaviour of rows of vortices distributed throughout two domains, that are separated by a twin boundary filled with strongly pinned vortices. A field difference is imposed across the TB, and the position of individual rows is determined by solving the London equation in an iterative process. **a**, Situation representing the one observed in Fig. 1c. The fields far from the TB differ by 0.2 T. **b**, Situation observed in Fig. 1d. The fields far from TB differ by 0.1 T. In both cases **a** and **b**, the current flowing along the TB is about $4 \times 10^8 \text{ A cm}^{-2}$, a value of the order of the depairing current for YBCO. The density of vortices along the TB is determined by the distance separating the TB from the first row of vortices in each domain. In both cases, this density is significantly higher than that observed within the domains.

at a field difference of only 0.2 T (Fig. 2a). To keep the first row to the left of the boundary at the observed position, the density of vortices in the boundary must be about twice the density in the rows away from it. The distance between the vortices trapped in the boundary is thus only about 12 nm, sufficiently small that in our image the vortices seem to overlap. In our simplified model the vortices form a square lattice, but this is not essential for the arguments given here. The behaviour described above is similar to the one that was developed by Ternovskii and Shekhata for surface barriers¹⁸. The calculated current density along the boundary for this configuration is found to be as high as $4 \times 10^8 \text{ A cm}^{-2}$. Although this value depends on our choice of λ , it will in any case be close to the depairing current.

In this picture, a large number of vortices will be pinned along the twin boundaries, which act as barriers, whereas the vortices inside the domains are relatively free to move. When the field varies significantly, some vortices will begin to leave the sample. Across a given boundary, one of the sides will see a faster field reduction than the other side because somehow vortices leak out faster. The vortices in the lower field region will gradually be pushed away from the boundary until the current along the twin boundary reaches its critical value. At this point flux lines can begin to cross from the other side and an extreme local critical state is established across the twin boundary. As long as flux lines leak out of the right-hand domain, we expect that the border between the vortex-free domain and the vortices to the right will be irregular and fluctuate in time, explaining the presence of the non-continuous lines in Fig. 1b. Note that in the investigated region, we did not find evidence for other pinning effects comparable to those produced by the boundary. When the field is increased to 6 T, the vortices enter from the right, leading to an increase of the local field in the domain. Correspondingly the vortices on the left-hand side are pushed away from the twinning boundary, but at a much smaller distance because the fields are now much higher. □

Received 25 February; accepted 9 September 1997.

- Blatter, G., Feigel'man, M. V. & Vinokur, V. M. Vortices in high-temperature superconductors. *Rev. Mod. Phys.* **66**, 1125–1388 (1994).
- Cubitt, R. *et al.* Direct observation of magnetic flux lattice melting and decomposition in the high T_c superconductor $\text{Bi}_{1.5}\text{Sr}_{1.95}\text{CaCu}_2\text{O}_{8+x}$. *Nature* **365**, 407–411 (1993).
- Schilling, A. *et al.* Calorimetric measurement of the latent heat of vortex-lattice melting in untwinned $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$. *Nature* **382**, 791–793 (1996).
- Roulin, M., Junod, A. & Walker, E. Flux line lattice melting transition in $\text{YBa}_2\text{Cu}_3\text{O}_{6.94}$ observed in specific heat experiments. *Science* **273**, 1210–1212 (1996).
- Junod, A. *et al.* Specific heat peaks observed at 16 T on the melting line of the vortex lattice in $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Physica C* **275**, 245–258 (1997).
- Gammel, P. L. *et al.* Observation of a novel vortex structure given by magnetic interactions near a sawtooth twin boundary in $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$. *Phys. Rev. Lett.* **69**, 3808–3811 (1992).
- Duran, C. A. *et al.* Real-time imaging of the magnetic flux distribution in superconducting $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$. *Nature* **357**, 474–477 (1992).
- Vlasov, V. K. *et al.* Study of the influence of individual twin boundaries on the magnetic flux penetration in $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$. *Phys. Rev. Lett.* **72**, 3246–3249 (1994).
- Welp, U. *et al.* Magneto-optical study of twin boundary pinning in $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$. *Physica C* **235–240**, 241–244 (1994).
- Hess, H. F., Robinson, R. B., Dynes, R. C., Valles, J. M. Jr. & Waszczak, J. V. Scanning-tunneling-microscope observation of the Abrikosov flux lattice and the density of states near and inside a fluxoid. *Phys. Rev. Lett.* **62**, 214–217 (1989).
- Renner, Ch., Kent, A. D., Niedermann, Ph., Fischer, Ø. & Lévy, F. Scanning tunneling spectroscopy of a vortex core from the clean to the dirty limit. *Phys. Rev. Lett.* **67**, 1650–1653 (1991).
- Maggio-Aprile, L., Renner, Ch., Erb, A., Walker, E. & Fischer, Ø. Direct vortex lattice imaging and tunneling spectroscopy of flux lines on $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$. *Phys. Rev. Lett.* **75**, 2754–2757 (1995).
- Erb, A., Walker, E. & Flükiger, R. BaZrO₃: The solution for the crucible corrosion problem during the single crystal growth of high- T_c superconductors $\text{REBa}_2\text{Cu}_3\text{O}_{7-x}$. *RE = Y, Pr. Physica C* **245**, 245–251 (1995).
- Renner, Ch., Niedermann, Ph., Kent, A. D. & Fischer, Ø. A versatile low-temperature scanning tunneling microscope. *J. Vac. Sci. Technol. A* **8**, 330–332 (1990).
- Kent, A. D., Renner, Ch., Niedermann, Ph. & Fischer, Ø. A low-temperature scanning tunneling microscope with in-situ sample cleaving. *Ultramicroscopy* **42–44**, 1631–1637 (1992).
- Crabtree, G. W. *et al.* Time-dependent Ginzburg-Landau simulations of vortex guidance by twin boundaries. *Physica C* **263**, 401–408 (1996).
- Keimer, B. *et al.* Vortex lattice symmetry and electronic structure in $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Phys. Rev. Lett.* **73**, 3459–3462 (1994).
- Ternovskii, F. F. & Shekhata, L. N. Structure of the mixed state near the boundary of a semi-finite type II superconductor of the second kind. *Sov. Phys.-JETP* **35**, 1202–1208 (1972).

Acknowledgements. We thank R. Flükiger, A. Takagi, J.-G. Bosch, M.G. Karkut and C.H. Ahn for experimental assistance and critical comments about the manuscript. This work was supported by the Swiss National Science Foundation.

Correspondence should be addressed to I.M.-A. (e-mail: Ivan.Maggio-Aprile@physics.unige.ch).

Femtosecond time-resolved X-ray diffraction from laser-heated organic films

Christian Rischel*, Antoine Rousse*, Ingo Uschmann†, Pierre-Antoine Albouy‡, Jean-Paul Geindre§, Patrick Audebert§, Jean-Claude Gauthier§, Eckhart Förster†, Jean-Louis Martin* & André Antonetti*

* Laboratoire d'Optique Appliquée, ENSTA—Ecole Polytechnique, CNRS URA1406, INSERM U451, Chemin de la Hunière, 91761 Palaiseau, France

† X-ray Optics Group, Institute of Optics and Quantum Electronics, Friedrich Schiller University Jena, Max-Wien-Platz 1, 07743 Jena, Germany

‡ Laboratoire de Physique des Solides, Bâtiment 510, Centre Universitaire, 91405 Orsay, France

§ Laboratoire pour l'Utilisation des Lasers Intenses, CNRS UMR100, Ecole Polytechnique, 91128 Palaiseau, France

The extension of time-resolved X-ray diffraction to the subpicosecond domain is an important challenge, as the nature of chemical reactions and phase transitions is determined by atomic motions on these timescales. An ultimate goal is to study the structure of transient states with a time resolution shorter than the typical period of vibration along a reaction coordinate (around 100 fs). Biological processes that can be initiated optically have been studied extensively by ultrafast infrared, visible and ultraviolet spectroscopy¹. But these techniques probe only electronic states, whereas time-resolved crystallography should be able to directly monitor atomic positions. Here we show that changes in the X-ray diffraction pattern from an organic film heated by a laser pulse can be monitored on a timescale of less than a picosecond. We have studied the response of a Langmuir–Blodgett multilayer film of cadmium arachidate to laser heating by observing changes in the intensity of one Bragg peak for different delays between the perturbing optical pulse and the X-ray probe pulse. A strong decrease in intensity is seen within a picosecond of heating, resulting from disorder introduced to the layers of cadmium atoms before thermal expansion of the film (which ultimately leads to its destruction) has time to occur.

Several techniques have emerged for generating an ultrafast X-ray probe synchronized with an optical trigger. Modern synchrotrons deliver intense polychromatic radiation in pulses with a duration of around 50 ps, and by gating the detection with a streak camera, time resolution down to 500 fs can be obtained; so far 1.5 ps resolution has been demonstrated². Much weaker pulses of 300 fs duration have been generated by Thomson scattering of intense laser pulses on a relativistic electron beam³. A third, completely different approach is to use the monochromatic $K\alpha$ X-ray flashes that are generated by suprathermal electrons when an ultrashort laser pulse is focused on a solid surface⁴. For very short plasma gradient scale lengths, the duration of the $K\alpha$ emission is governed by the thermalization time of the electrons in the bulk of the target and by the duration of the laser pulse^{5–7}. Thus, theory indicates that X-ray pulses below 100 fs might be obtained from femtosecond laser pulses with this technique. Using toroidally bent single crystals, the X-rays can be collected over a relatively large solid angle and focused onto the sample without significant increase in the pulse length⁸.

We have used the X-ray $K\alpha$ -emission from a laser-irradiated silicon target to probe the multilayer film sample heated by an ultrashort laser pulse. A fairly long time (probably hundreds of picoseconds) after the heating laser pulse has reached the sample, the film completely evaporates, and the experiment follows the development of the layered structure in the time between heating and evaporation. To obtain full information about the atomic

structure, a large number of Bragg reflections would need to be monitored. In this study, we record only the first-order reflection from which we can deduce the distance between the layers and follow how the initial confinement of the cadmium atoms to narrow planes is disrupted.

A sketch of the set-up is shown in Fig. 1 (see Methods section for details). The experiment measures the diffracted intensity as a function of angle (the rocking curve), either the reference rocking curve $R(\theta)$ without heating or the test $T(\theta)$ with some delay between the X-ray and heating pulses. The upper panel in Fig. 2 shows $R(\theta)$. The changes in the diffracted signal due to the heating are most clearly visualized by the transient signal $T(\theta) - R(\theta)$, which is shown in the lower panel in Fig. 2 for selected delay times. As expected, there is no transient signal at negative delay times (X-ray pulse arriving before laser pulse), but for the earliest positive delays the signal drops visibly, and after 1 ps no further changes can be detected, given the signal-to-noise ratio of the present experiment. For a monocrystalline sample, the width of the rocking curve is inversely proportional to the length over which the lattice layers give similar contributions to the scattering amplitude. This length can be affected by absorption as well as by extinction caused by diffraction, but in the present case it is determined by the vertical correlation length of 1,100 Å characterizing the perfection of the layered structure. Earlier picosecond to nanosecond experiments on laser-heated crystals of silicon⁹, gold¹⁰ or gallium arsenide¹¹ all show significant broadening of the initially narrow rocking curves. This indicates that the heating is inhomogeneous over the penetration depth of the X-rays, which is several micrometres for these crystals. In some cases one observes a shift of the Bragg angle, reflecting a change in the spacing between lattice planes. These effects are not visible in our data, because the film is modified homogeneously over the short correlation length, and because the diffraction is rapidly extinguished by thorough dis-

order inside each layer of cadmium atoms (see below) before bulk expansion or compression of the sample can occur.

The decrease in diffracted intensity as function of time is shown in Fig. 3, normalized to correct for the imperfect overlap of the X-rays and the heating pulse. The signal from the heated film is seen to drop rapidly by 75%. The solid line shows an exponential curve with a rise-time of 600 fs, representing an upper limit to the time constant. The data do not permit a lower limit to be set. As the experimental curve is the convolution of the material response with the probing X-ray pulse, both the material response of the sample to the heating and the duration of the $K\alpha$ X-ray pulse are included in 600 fs. The theory strongly supports an X-ray pulse duration very close to 100 fs (see above). Under this reasonable assumption, the results shown in Fig. 3 actually correspond to the description of the material response. The signal remains constant out to 150 ps, so the disappearance of the last 25%, probably due to evaporation of the film, must take place on a slower timescale.

The loss of diffracted intensity on heating of the film indicates a displacement of the cadmium atoms out of the layers. As most of the electrons follow the nuclear movements, departures from the equilibrium positions lead to a broader average electronic distribution. Quantitatively, one may set the electronic density f to a gaussian function of the distance z to the initial centre of the Cd layers, $f(z) \propto \exp(-\frac{1}{2}[z/\sigma(t)]^2)$; the intensity of the Bragg peak is then well approximated by

$$I(t) \approx I_0 \exp \left[- \left(\frac{2\pi\sigma(t)}{d} \right)^2 \right]$$

where $\sigma(t)$ is the half-width of the electron distribution, d is the distance between the cadmium layers and I_0 is the intensity for an infinitely sharp distribution ($\sigma = 0$). For the cold film $\sigma \approx 1.5$ Å, giving $I = 0.97I_0$, and it is seen that the intensity has dropped to

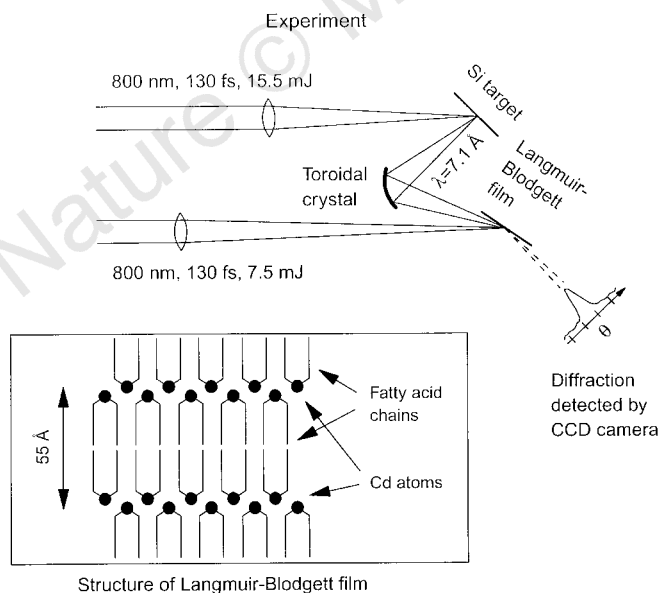


Figure 1 Experimental set-up. A laser pulse is focused on a Si target to create a flash of $K\alpha$ -photons, which are collected and refocused by a toroidal crystal. It has been shown theoretically⁹ that the diffraction of the crystal used broadens the X-ray pulse by less than 50 fs. The sample is 2,300-Å-thick Langmuir-Blodgett film of 83 layers of cadmium arachidate (inset) which is heated by a second laser pulse with some delay relative to the X-ray pulse. As this beam is not parallel to the X-rays, the delay between the two varies over the heated zone, resulting in a temporal mismatch of 120 fs. The Bragg diffraction is due to the high electronic density in the layers of Cd atoms, with virtually no contribution from the fatty acid chains. Both Si target and film are replaced after each shot.

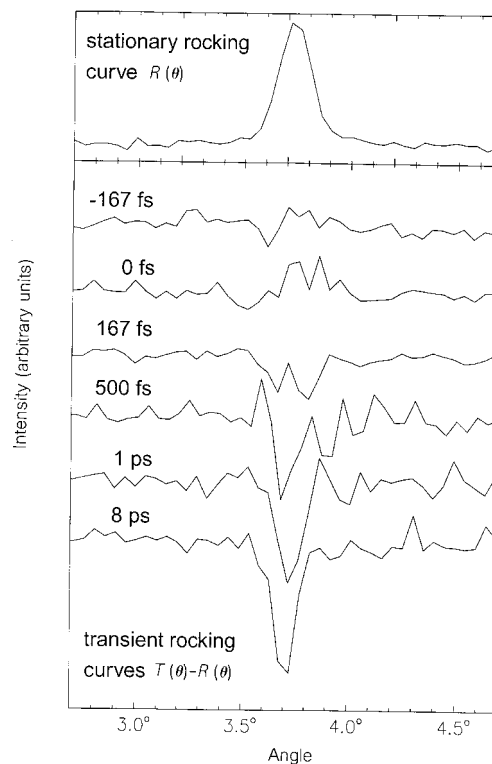


Figure 2 Stationary rocking curve of the sample (upper panel) and transient rocking curves for selected delays between X-ray and heating pulses (lower panel). Each transient curve is an average of 10 to 30 shots.

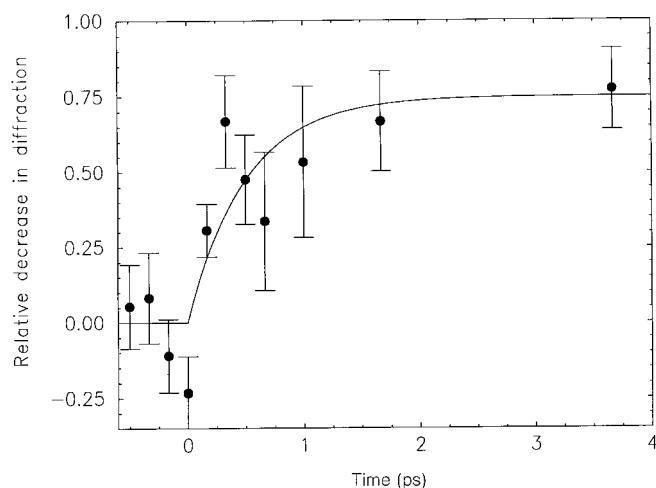


Figure 3 Decrease in diffracted intensity, calculated as $-\Sigma_{\theta}[I(\theta) - R(\theta)]R(\theta)/\Sigma_{\theta}R(\theta)^2$, which weights the contribution of each point by the intensity of the reference. The values are normalized to the value measured after full ablation of the film, to correct for the fact that the X-ray focus is larger than the laser heated spot on the film. A value of 0 corresponds to the diffraction from the unperturbed film, a value of 1 corresponds to signal measured long (several minutes) after the heating laser pulse. Error bars represent ± 1 standard deviation.

$I_0/e = 0.37I_0$ when σ reaches $d/2\pi = 8.8 \text{ \AA}$. The Debye–Waller theory¹² describes X-ray diffraction from heated crystals by applying similar considerations to account for the effects of increased harmonic lattice vibrations. In our experiments, the sample is superheated high above the melting point, so the Cd atoms move far from their initial positions, but the effect of the broader electron distribution is still to reduce the reflectivity of the layered structure.

In the subpicosecond pulse regime, laser energy is deposited in dielectric materials by multiphoton ionization followed by collisional electron avalanche up to the critical electron density n_c (in the present case $1.7 \times 10^{21} \text{ cm}^{-3}$), where breakdown damage¹³ and supercritical plasma formation is induced. For our laser intensity, multiphoton ionization rates¹⁴ for cadmium, oxygen and carbon atoms predict single ionization (electronic density larger than n_c) during the rising edge of the short optical pulse, and collisional heating of the electrons occurs before any significant transfer of energy to the lattice. A large number of covalent bonds are broken, and as the atoms find themselves far from equilibrium on the strongly modified potential energy surface, the electronic energy is rapidly coupled into nuclear motion. Motion of cadmium atoms out of the layers now leads to a reduced diffracted intensity as explained above. A primitive estimate of the time required for broadening of the distribution is $d/(2\pi v) = 300 \text{ fs}$, where $v = 28 \text{ \AA ps}^{-1}$ is the mean velocity of cadmium atoms at 5 eV, in rough agreement with the experimental time constants.

Following the recent demonstration of X-ray absorption probing a chemical reaction with picosecond time resolution¹⁵, our work represents further progress in ultrafast characterization of reaction dynamics by time-resolved X-ray diffraction. The discovery that the diffraction of thin films is switched off ultrarapidly confirms that large structural rearrangements can occur on the femtosecond timescale in condensed matter. □

Methods

The experiments were done with a 10 Hz laser delivering 130 fs pulses at 800 nm in two beams with a total energy of 23 mJ at the entry to the vacuum chamber. A pulse of energy 15.5 mJ was passed through a pre-pulse system which advances the central 10% of the beam 9 ps with respect to the rest of the pulse, and then focused on a silicon target to generate the X-ray flash. The initial pulse serves to generate a pre-plasma, which converts the laser pulse energy into K α -photons more efficiently than a solid¹⁶. In our experiment, this technique increased the X-ray yield by a factor of 7. The scale length of the electron density gradient remains below 0.2 times the laser wavelength¹⁶, which ensures short K α -emission⁷. The pre-pulse alone does not cause emission of K α -photons.

The emitted Si K α photons at 7.13 Å were collected over a solid angle of 5×10^{-3} steradian by a toroidally bent crystal of quartz with 100 orientation

and focused onto a 300- μm -high vertical line on the film. The focus contained 5×10^4 K α photons per shot. Because of the relatively low Bragg angle of 3.7° the X-ray focus was stretched to 1.25 mm in the horizontal direction. The diffracted signal containing about 50 photons per shot was detected by a Photometrics CCD camera positioned at 5 cm from the focus. As the toroidal crystal delivers nearly monochromatic radiation over a horizontal aperture angle that is much larger than the angular width of the film's rocking curve, the diffracted intensity as a function of the angle directly gives this curve^{10,11}, broadened by the horizontal extension of the X-ray focus. A laser pulse of 7.5 mJ with a spot diameter of 250 μm was used to heat the film, with an incident angle of 18° with respect to the target surface. As this spot does not entirely cover the X-ray focus, the noise on the normalized intensity of Fig. 3 is larger than expected from simple counting statistics. Before each shot the diffracted intensity from the cold film was probed with 5 shots without a heating pulse. One sequence of reference shots and 'live' shot takes about 30 seconds.

Received 19 June; accepted 17 September 1997.

- Vos, M. H. & Martin, J.-L. Femtosecond biology. *Annu. Rev. Biophys. Biomol. Struct.* **21**, 199–222 (1992).
- Larsson, J. *et al.* Time-resolved X-ray diffraction from laser-excited crystals. in *Application of High Field & Short Wavelength Sources* (Plenum, New York, in the press).
- Schoenlein, R. W. *et al.* Femtosecond X-ray pulses at 0.4 Å generated by 90° Thomson scattering: a tool for probing the structural dynamics of materials. *Science* **274**, 236–238 (1996).
- Rousse, A. *et al.* Efficient K α X-ray source from femtosecond laser-produced plasmas. *Phys. Rev. E* **50**, 2200–2207 (1994).
- Murnane, M. M., Kapteyn, H. C., Rosen, M. D. & Falcone, R. W. Ultrafast X-ray pulses from laser-produced plasmas. *Science* **251**, 531–536 (1991).
- Kieffer, J. C. *et al.* Ultrafast X-ray sources. *Phys. Fluids B* **5**, 2676–2681 (1993).
- Geindre, J. P. *et al.* X-rays generated by femtosecond laser-produced plasmas. in *Ultrafast Phenomena VIII* (Springer, Berlin, 1993).
- Chukhovskii, F. N. & Förster, E. Time-dependent X-ray Bragg diffraction. *Acta Crystallogr.* **A51**, 668–672 (1995).
- Larson, B. C., White, C. W., Noggle, T. S., Barhorst, J. F. & Mills, D. M. Time-resolved X-ray diffraction measurement of the temperature and temperature gradients in silicon during pulsed laser annealing. *Appl. Phys. Lett.* **42**, 282–284 (1983).
- Chen, P., Tomov, V. & Rentzepis, P. M. Time resolved heat propagation in a gold crystal by means of picosecond X-ray diffraction. *J. Chem. Phys.* **104**, 10001–10007 (1996).
- Rose-Petrick, C. *et al.* Measurements of ultrafast dynamics in GaAs crystals using time-resolved X-ray diffraction. in *Application of High Field & Short Wavelength Sources* (Plenum, New York, in the press).
- Compton, A. H. & Allison, S. K. X-rays in Theory and Experiment. (van Nostrand, Princeton, New Jersey, 1954).
- Stuart, B. C. *et al.* Nanosecond-to-femtosecond laser-induced breakdown in dielectrics. *Phys. Rev. B* **53**, 1749–1761 (1996).
- Amosov, M. V., Delone, N. B. & Krainov, V. P. Tunnel ionization of complex atoms and of atomic ions in an alternating electromagnetic field. *Sov. Phys. JETP* **64**, 1191–1194 (1986).
- Räksä, F. *et al.* Ultrafast X-ray absorption probing of a chemical reaction. *J. Chem. Phys.* **104**, 6066–6069 (1996).
- Bastiani, S. *et al.* Experimental study of the interaction to subpicosecond laser pulses with solid targets of varying initial scale lengths. *Phys. Rev. E* (in the press).

Acknowledgements. We gratefully acknowledge the support of the laser staff at Laboratoire d'Optique Appliquée where the experiments were carried out. C.R. was supported by a grant from the Carlsberg Foundation. This work was supported by the Centre National de la Recherche Scientifique and by the European Community under the Large Facility Program.

Correspondence should be addressed to A.R. (e-mail: rousse@ensta.ensta.fr).

Observation of magneto-chiral dichroism

G. L. J. A. Rikken & E. Raupach

Grenoble High Magnetic Field Laboratory, Max Planck Institut für Festkörperforschung/CNRS, BP 166, F-38042 Grenoble, France

Arago's discovery in 1811 of natural optical activity in chiral crystals and Faraday's discovery in 1846 of magnetically induced optical activity have contributed much to our understanding of the wave nature of light and the electronic properties of molecules. Both effects are manifest as a rotation in the polarization of transmitted light: the former is due to the intrinsic properties of media that lack mirror symmetry, whereas the latter (which occurs in all materials) is due to magnetic-field-induced changes in the optical properties. The apparent similarity of these two effects motivated Pasteur to search in vain for a link between the two phenomena¹. Such a link—which can be regarded as arising either from a magnetically induced change of natural optical activity or from the difference in magnetic optical activity of the two enantiomers of a chiral medium—has been predicted to exist^{2–7}, although it is expected to be very weak. Here we report the experimental observation of this 'magneto-chiral' optical effect and a demonstration of its enantioselectivity. The existence of this effect may be important in the context of fundamental interactions between light and matter, and in molecular spectroscopy.

There is indeed a strong phenomenological resemblance between natural and magnetic optical activity. Both represent a difference in absorption and refraction between left and right circularly polarized light, the former in chiral media, the latter in media subject to a magnetic field **B** parallel to the wavevector of the light **k**. The two effects, however, have completely different origins. Natural optical activity (NOA) is a result of a nonlocal optical response in media that lack mirror symmetry, whereas magnetic optical activity (MOA) results from the breaking of time-reversal symmetry by a

magnetic field. Under conditions where both symmetries are broken simultaneously, a physical situation exists that allows for an additional optical effect. The existence of such a cross-effect between natural and magnetic optical activity was first implied by calculations addressing magneto-optical rotation by molecules². On the basis of symmetry arguments, Baranova *et al.* inferred the existence of a contribution to the dielectric constant of a chiral medium subject to a magnetic field, of the form **B**·**k**, independent of polarization, and of opposite sign for the two enantiomers^{3,4}. Independently, Wagnière and Meier calculated a magnetically induced shift in the absorption and emission rates of chiral molecules^{5,6}. Barron and Vbrancich have unified these results using a molecular theory⁷, and showed that the optical properties of chiral systems are different for light propagating parallel or antiparallel to the magnetic field direction. This difference is independent of the polarization state of the light, and has opposite sign for the two enantiomers. In analogy with natural and magnetic optical activity, they named the effect magneto-chiral dichroism (MChD) in absorption or emission, and magneto-chiral birefringence (MChB) in refraction. Related magnetochromic effects have also been considered⁸.

Although there seems to be theoretical unanimity on the existence of magneto-chiral anisotropy, it has never been observed experimentally. This is no doubt due to the expected weakness of the effect, as both MOA and NOA are in most materials only small perturbations of the optical properties and their cross-effect is therefore likely to be even smaller. The most strongly chiral transitions reported^{9,10} are the ⁵D₀ → ⁷F_{1,2} luminescent transitions in tris(3-trifluoroacetyl-±-camphorato) europium(III) complexes, where ± indicates handedness (Eu((±)tfc)₃; see Fig. 1). These transitions also show considerable magnetic circular dichroism for the ion in aqueous solution¹¹. Such complexes are therefore likely candidates to show a significant magneto-chiral effect⁷. The experiment that we have done to observe MChD is essentially that suggested by Wagnière⁶ (see Fig. 2). It measures the difference in luminescence intensity in the directions parallel and antiparallel to the externally applied magnetic field **B**. To increase sensitivity, the

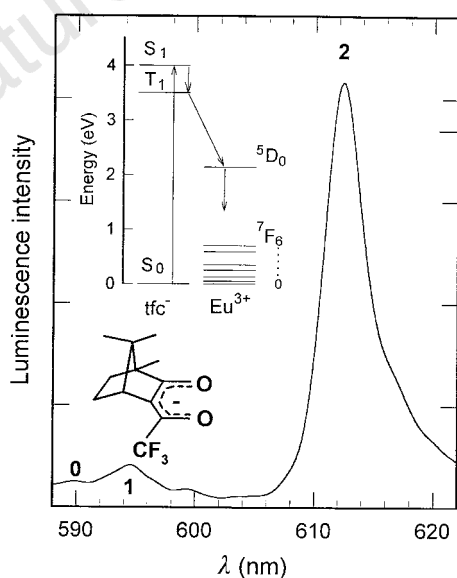


Figure 1 Luminescence of the Eu(±)tfc₃ complex dissolved in deuterated dimethyl sulphoxide, excited around 350 nm. Peaks are labelled with the *J* subscript of the final ⁷F_{*J*} level. Inset depicts the molecular structure of the Eu(tfc)₃ complex and shows the relevant energy levels, excitation, transfer and luminescence processes.

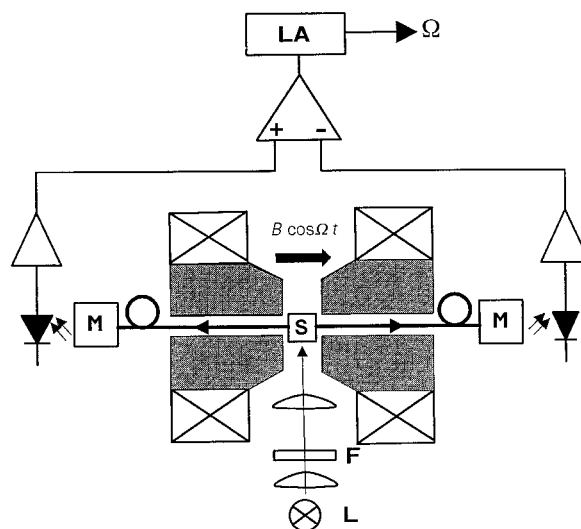


Figure 2 Schematic set-up for detecting magneto-chiral luminescence anisotropy. Sample (S) is excited around 350 nm with filtered (F), unpolarized light from a mercury-discharge lamp (L). Luminescence is collected in the directions parallel and antiparallel to the magnetic field by optical fibres (numerical aperture 0.47), specific emission wavelengths are selected by a grating monochromator (M, spectral resolution ~2 nm), and detected by silicon photodiodes. The intensity difference between the two directions is phase-sensitively detected by a lock-in amplifier (LA) at the frequency $\Omega = 0.9$ Hz at which the magnetic field is alternated.

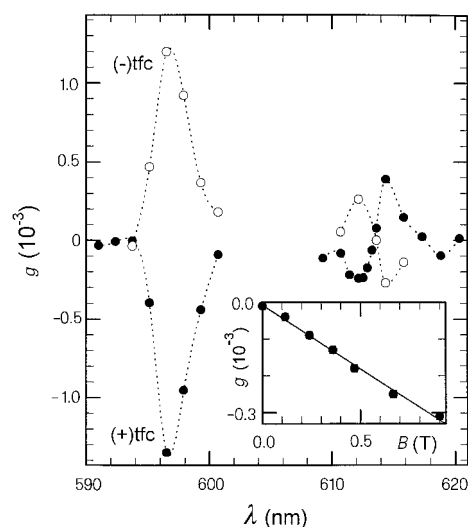


Figure 3 Magneto-chiral luminescence anisotropy of $\text{Eu}(\pm)\text{tfc}_3$ complexes, dissolved in deuterated dimethyl sulphoxide (5% wt/wt), as a function of luminescence wavelength, with excitation around 350 nm. The (alternating) magnetic field strength was 0.9 T. Filled symbols represent the $\text{Eu}(+)\text{tfc}_3$ complex, open symbols its enantiomer, the $\text{Eu}(-)\text{tfc}_3$ complex. Dashed lines are only meant to guide the eye. The inset shows the magneto-chiral anisotropy of the $\text{Eu}(-)\text{tfc}_3$ complex as a function of magnetic field strength, excitation around 350 nm, luminescence detected at 615.8 nm. Solid line is a linear fit.

magnetic field is alternated and the intensity difference is measured with a phase-sensitive detection method. Factors related to excitation intensity, complex concentration and sample geometry are eliminated by dividing the output of the lock-in amplifier by the total, static luminescence signal. The resulting quantity is the magneto-chiral anisotropy factor g , given by

$$g = \frac{\frac{\partial}{\partial B}(I_{B\parallel k} - I_{B\perp k})B}{I_{B\parallel k} + I_{B\perp k}} \quad (1)$$

where subscript $B\parallel k$ means B parallel to k , and subscript $B\perp k$ means B antiparallel to k . The use of fibres with a finite numerical aperture implies that some light that is not propagating either perfectly parallel or antiparallel to the magnetic field is collected. The measured value of g will therefore be smaller than the definition above.

Figure 3 shows the experimental results for g for the two enantiomers of the complex, confirming significant magneto-chiral anisotropy at the two transitions $^5D_0 \rightarrow ^7F_1$ and $^5D_0 \rightarrow ^7F_2$. An essential characteristic of MChD is that g should be of opposite sign for the two enantiomers. This feature is observed within the experimental accuracy. The inset shows the expected linear magnetic field dependence of g . Figure 3 therefore demonstrates the existence of magneto-chiral dichroism and its enantioselectivity. Although it is clear that the magneto-chiral effect is not just the product of natural and magnetic circular dichroism (NCD and MCD), one could intuitively expect as a first estimate for the strength of MChD something proportional to the product of the strengths of NCD and MCD. This is supported by Baranova *et al.*⁴ who consider MOA as an additional dispersive effect, acting both on the normal dielectric constant and on the NOA. Oversimplified as this view may be, it is capable of predicting the proper symmetry and polarization independence of the magneto-chiral effect. It predicts the anisotropy factor of MChD to be equal to the product of the dissymmetry factors of NCD and MCD. In our case, for the $^5D_0 \rightarrow ^7F_1$ transition this would amount to $g \approx 5 \times 10^{-3}$ and for the $^5D_0 \rightarrow ^7F_2$ transition to $g \approx 4 \times 10^{-4}$ (refs 9, 11). It should be noted that there is some discussion on the exact value of the NCD dissymmetry value of these complexes,^{9,10} and that the MCD values were obtained on aqueous Eu^{3+} ions. The agreement of the above estimate with what we observe experimentally is reasonable in view of the crudeness of the model and the uncertainty in the parameters.

The $^5D_0 \rightarrow ^7F_1$ and the $^5D_0 \rightarrow ^7F_2$ transitions have clearly different MChD, both in strength and lineshape. This is not surprising as the first is a magnetic dipole allowed transition for the free ion, with a very strong NCD dissymmetry in the tfc complex, and the second is an electric dipole allowed transition for the free ion¹². The theory of Barron and Vrbancich shows that MChD has many possible

contributions whose lineshapes may be proportional to either the emission lineshape or to the product of this lineshape with its own derivative. The Eu^{3+} complexes we have studied qualify to show both types of contribution, their relative strengths depending on a delicate interplay between electric dipole, magnetic dipole and electric quadrupole transition moments. For the $^5D_0 \rightarrow ^7F_2$ transition we observe for the MChD a derivative-type lineshape, which one would also expect on the basis of the simple model of Baranova *et al.*⁴. For the $^5D_0 \rightarrow ^7F_1$ transition the MChD lineshape resembles most the emission lineshape, although its maximum is offset from the emission maximum. It may therefore still be derivative-like, with its short-wavelength part heavily distorted by the presence of the $^5D_0 \rightarrow ^7F_0$ transition. A quantitative comparison between the theory of Barron and Vrbancich and our experiments must await the experimental determination of all the transition moments involved.

Now that magneto-chiral anisotropy has been shown to exist in emission (that is, in the imaginary part of the refractive index) the relations between Einstein's coefficients in radiative processes automatically imply its existence in absorption. Furthermore, the Kramers–Kronig relations imply its existence in the real part of the refractive index, and thus in refraction (MChB). □

Received 6 May; accepted 30 September 1997.

1. Mason, S. F. in *Circular Dichroism* (eds Nakanishi, K., Berova, N. & Woody, R. W.) (VCH, New York, 1994).
2. Groenewege, M. P. A theory of magneto-optical rotation in diamagnetic molecules of low symmetry. *Molec. Phys.* **5**, 541–563 (1962).
3. Baranova, N. B., Bogdanov, Yu. V. & Zeldovich, B. Ya. Electrical analog of the Faraday effect and other new optical effects in liquids. *Opt. Commun.* **22**, 243–247 (1977).
4. Baranova, N. B. & Zeldovich, B. Ya. Theory of a new linear magnetorefractive effect in liquids. *Molec. Phys.* **38**, 1085–1098 (1979).
5. Wagnière, G. & Meier, A. The influence of a static magnetic field on the absorption coefficient of a chiral molecule. *Chem. Phys. Lett.* **93**, 78–81 (1982).
6. Wagnière, G. Magneto-chiral dichroism in emission. Photoselection and the polarization of transitions. *Chem. Phys. Lett.* **110**, 546–551 (1984).
7. Barron, L. D. & Vrbancich, J. Magneto-chiral birefringence and dichroism. *Molec. Phys.* **51**, 715–730 (1984).
8. Andrews, D. L. & Bittner, A. M. Influence of a magnetic field on line intensities in the optical spectra of free molecules. *J. Chem. Soc. Faraday Trans.* **87**, 513–516 (1991).
9. Brittain, H. G. & Richardson, F. S. Circularly polarized emission studies on the chiral nuclear magnetic resonance lanthanide shift reagent tris(3-trifluoroacetyl-d-camphorato)europium(III). *J. Am. Chem. Soc.* **98**, 5858–5863 (1976).
10. Schippers, P. H., van den Beukel, A. & Dekkers, H. P. J. M. An accurate digital instrument for the measurement of circular polarisation of luminescence. *J. Phys. E: Sci. Instr.* **15**, 945–950 (1982).
11. Görller-Walrand, C. & Godemont, J. MCD of the Eu^{3+} ion in aqueous solution. Analysis of the $^5D_{0,1,2} \leftarrow ^7F_{0,1,2}$ transitions. *J. Chem. Phys.* **67**, 3655–3658 (1977).
12. Freed, S. & Weissman, S. I. Multipole nature of elementary sources of radiation—Wide angle interference. *Phys. Rev.* **60**, 440–442 (1941).

Acknowledgements. We thank H. Krath for technical assistance, E. W. Meijer and B. Voss for providing samples, and P. Wyder, G. Martinez, L. Jansen, A. Sparenberg and B. van Tiggele for critical reading of the manuscript. The Grenoble High Magnetic Field Laboratory is a 'laboratoire conventionné aux universités UJF et INP de Grenoble'.

Correspondence should be addressed to G.L.J.A.R. (e-mail: rikken@labs.polycnrs-gre.fr).

Complex pathways in dissociative adsorption of oxygen on platinum

T. Zambelli, J. V. Barth*, J. Wintterlin & G. Ertl

Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4–6,
D-14195 Berlin, Germany

Gas adsorption on solid surfaces is the basis of heterogeneous catalysis. Gas–surface interactions may be complex and in many cases the fundamental mechanisms of the chemisorption process are hard to discern. The macroscopic kinetics of a heterogeneous catalytic reaction are usually modelled within the Langmuir model¹, which assumes that free adsorption sites are occupied at random. The adsorption of oxygen on a platinum (111) surface has been studied extensively as a model system for surface chemical processes generally^{2–15}, owing to its significance in platinum catalysed oxidation reactions such as that of CO and NO. Here we show that even for this well studied system the chemisorption process may be much more complicated than the Langmuir model implies. Our observations with the scanning tunnelling microscope show that the dissociation probability for an oxygen molecule becomes affected by chemisorbed species in the vicinity that have dissociated already. This introduces a dynamic heterogeneity in the adsorption mechanism which leads to kinetically limited ordering of the adsorbate. This effect is likely to be quite general and to affect the bulk kinetics of catalytic reactions conducted at the high temperatures and pressures of most industrial heterogeneous catalysis.

It is believed that, in the regime of low incident kinetic energy, the dissociative adsorption of oxygen on Pt(111) proceeds by sequential population of precursor states^{10–12}—that is, a physisorbed and a chemisorbed molecular oxygen species, which can be stabilized on the surface if the crystal temperature is reduced to values below 30 K and 160 K, respectively. We have used the scanning tunnelling microscope (STM) to study the initial distributions of chemisorbed oxygen on Pt(111) resulting from interaction with O₂ between 50 K and 160 K. The experiments were done in an ultrahigh-vacuum chamber (base pressure 3×10^{-11} torr) with a home-built STM cooled by liquid helium. Sample preparation and characterization followed procedures described previously¹⁴. The STM data were obtained in the constant-current mode with typical tunnelling currents of 0.1–1 nA and voltages of 30–500 mV (the possible effect of tip-induced dissociation of molecules reported recently¹⁶ was carefully investigated and can be excluded). Samples were exposed to oxygen by backfilling the chamber. In agreement with electron energy-loss spectroscopy⁵ and recent STM observations¹⁶, we find that for small coverages oxygen adsorbs dissociatively down to about 95 K. At even lower temperatures adsorbed O₂ molecules may accumulate on the surface.

Thermally activated surface diffusion of chemisorbed O atoms is suppressed on the timescale of the experiments at temperatures below about 160 K (ref. 14) and hence the STM images reproduced in Fig. 1, recorded after exposing the Pt(111) surface to 1 L O₂ (1 L = 10^{-6} torr s) at sample temperatures between 50 and 160 K, reflect the distributions of adsorbed particles just after thermal accommodation following equilibration. At the highest temperature, $T = 160$ K, the O-atom coverage corresponds to only about 0.016 monolayers, from which an initial dissociative sticking coefficient (the probability that an impinging oxygen gas molecule is

dissociatively chemisorbed on the clean Pt surface) of $S_0 = 0.03$ is derived (1 monolayer corresponds to 1 O atom per Pt substrate atom), in agreement with earlier observations¹². The adatoms always appear as pairs (imaged as dark spots) which are randomly distributed across the surface. In agreement with previous findings, atoms in these pairs have, on the average, a separation of twice the lattice constant of the Pt surface, $2a = 5.6$ Å (ref. 14). There is no preferential adsorption on surface defects, such as atomic steps. On lowering the substrate temperature to 115 K, we find that cluster formation of the oxygen pairs is predominant, where most (~70%) of the atoms are assembled in small islands, corresponding to a marked increase of the total coverage to 0.022 monolayers O. At $T = 110$ K, these clusters grow quasi-one-dimensionally with lengths between 10 Å and 50 Å. Atomic resolution data reveal that they grow preferentially along the close-packed directions of the surface; accordingly branched clusters form mostly 120° angles reflecting the three-fold symmetry of the substrate. At 105 K the atoms are almost exclusively arranged in such chains, which are partly interconnected forming an irregular network. A change in the island morphology occurs on reducing the temperature below 100 K: triangular oxygen clusters preferentially evolve at the points where three differently oriented chains come together. With further temperature reduction the island distribution becomes more inhomogeneous with the coexistence of wide patches

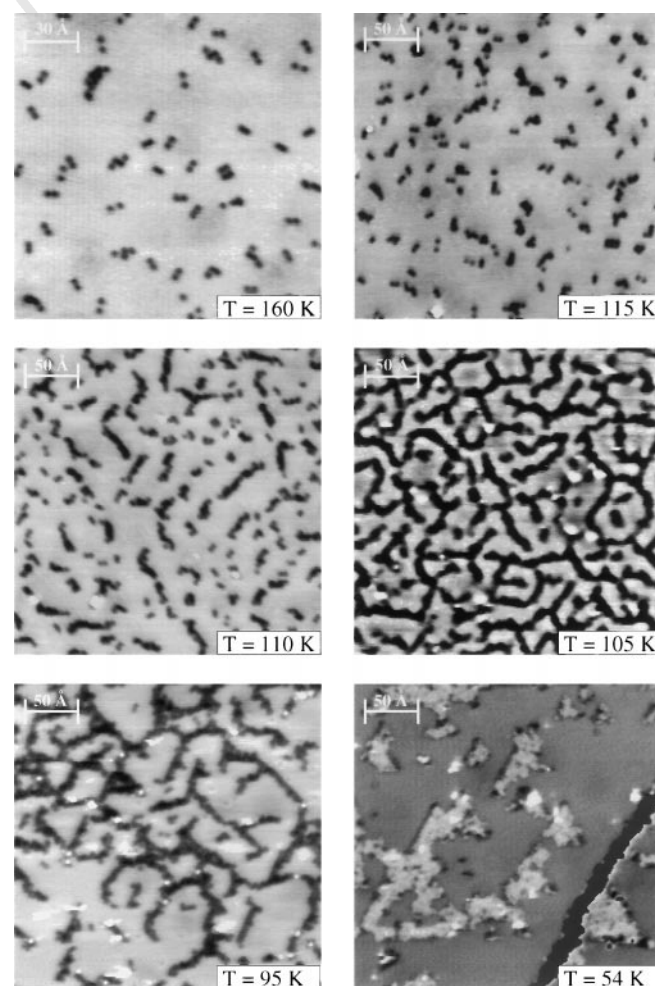


Figure 1 Temperature dependence of oxygen adsorption on Pt. These STM images of Pt(111) were recorded after adsorption of 1 L of O₂ at the indicated temperatures, and reveal the temperature-dependent sticking and anisotropic growth of oxygen islands on the surface. Scale bar in top left image, 30 Å; scale bar in all other images, 50 Å.

* Permanent address: Institut de Physique Expérimentale, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland.

of uncovered Pt and oxygen clusters, and the chains are continuously replaced by the more compact triangular islands. Finally, for temperatures below 80 K formation of one-dimensional structure ceases and exclusively compact triangular islands exist. In contrast to the atom islands formed at higher temperatures they are always imaged as protrusions on the Pt substrate. They are thus associated with molecularly adsorbed oxygen, in agreement with recent other STM observations¹⁶.

At first sight, one is tempted to attribute the formation of adsorbed oxygen (O_{ad}) chains to a locally increased dissociative sticking coefficient for gas molecules that impinge near already chemisorbed species, as was suggested to explain the increasing sticking coefficient with increasing coverage at low temperatures^{5,12}. This idea can be rejected if one analyses STM data for very small coverages in the earliest stages of oxygen uptake. Such data reveal that the fraction of clusters composed of four or more oxygen atoms far exceeds the maximum value to be expected from this effect. Furthermore, the strong temperature dependence of both the

sticking coefficient and the shape of the oxygen agglomerates formed cannot be explained this way. The observations, however, can be entirely rationalized by the assumption of a mobile adsorbed molecular precursor state. The mechanism of island formation is attributed to the action of already chemisorbed O atoms as active sites for trapping and dissociation of this precursor.

The temperature dependence of the sticking is thus understood as the result of the competition between precursor mobility, allowing it to reach the active sites for equilibration, its spontaneous dissociation on the clean surface and thermal desorption of the molecules from their weakly bound precursor state. With decreasing temperature the lifetime τ of the precursor at the surface increases. But as the activation energy for surface diffusion is generally only a fraction of the adsorption energy, this causes an increase in the mean free path of the precursor molecule on the surface (this is just the opposite behaviour to non-evaporating particles, where diffusion lengths increase with the temperature). The resulting higher probability of reaching an active site (an O atom) accounts for the marked increase of the dissociative sticking coefficient below 120 K reported previously¹² and also observed here. An estimate of the mobility and adsorption energy criteria required for agglomeration, using the known energy of the physisorption potential (~ 100 meV; ref. 15), indicates that the mobility in the physisorbed molecular state is the essential parameter for this behaviour.

Our data suggest interactions between chemisorbed O atoms and precursor molecules whereby the dissociation probability of the latter is locally increased. This mechanism is substantiated with the help of the atomic-resolution STM data reproduced in Fig. 2. Whereas at 160 K essentially isolated pairs of O atoms are formed, at 140 K small agglomerates are generally found (Fig. 2a), whose average size increases as the temperature is lowered further. The atomic-resolution image at 105 K (Fig. 2b) in addition resolves the growth of chain-like configurations of O_{ad} -pairs, preferentially following the close-packed directions of the substrate. This quasi-one-dimensional growth indicates that the effect of the O_{ad} -atoms on the precursors has an anisotropic nature, which presumably results from the local arrangement of the oxygen atoms within the agglomerates. We suggest that the dissociation probability of the O_2 -precursors is higher at the protruding chain ends than near the adatoms within a chain which are more highly coordinated with neighbouring atoms and consequently have a different electronic environment. The resulting anisotropic growth is sketched in the model in Fig. 2. A simple hit-and-stick mechanism would cause the formation of fractal islands, as found for example in two-dimensional simulations of diffusion-limited aggregation¹⁷, in contrast to the present observations. On oxygen adsorption at even lower temperatures (below ~ 100 K) the transition to compact island formation sets in until for $T < 80$ K exclusively compact molecular islands of the type shown in Fig. 1 (for $T = 54$ K) evolve. For their formation, the precursors must be sufficiently mobile along the island edges to ensure the observed triangular shapes rather than ramified growth. These observations signal that the interactions between the precursors and adsorbed molecules or atoms are attractive.

Our data demonstrate the complexity of an apparently simple surface reaction, the precursor-mediated dissociative adsorption of a molecule on a metal surface, resulting from the precursor mobility and local reactivity of chemisorbed particles, and the interaction anisotropy between the adsorbed species and the precursor. This effect is essentially kinetic in nature and may lead to metastable configurations of the adsorbed particles. It cannot be treated theoretically simply by including interaction energies between adsorbed particles in statistical concepts leading to equilibrium properties. Any catalytic reaction is governed by the same kinds of elementary processes, even if these become faster at the higher pressures and temperatures of 'real' conditions. The effects are

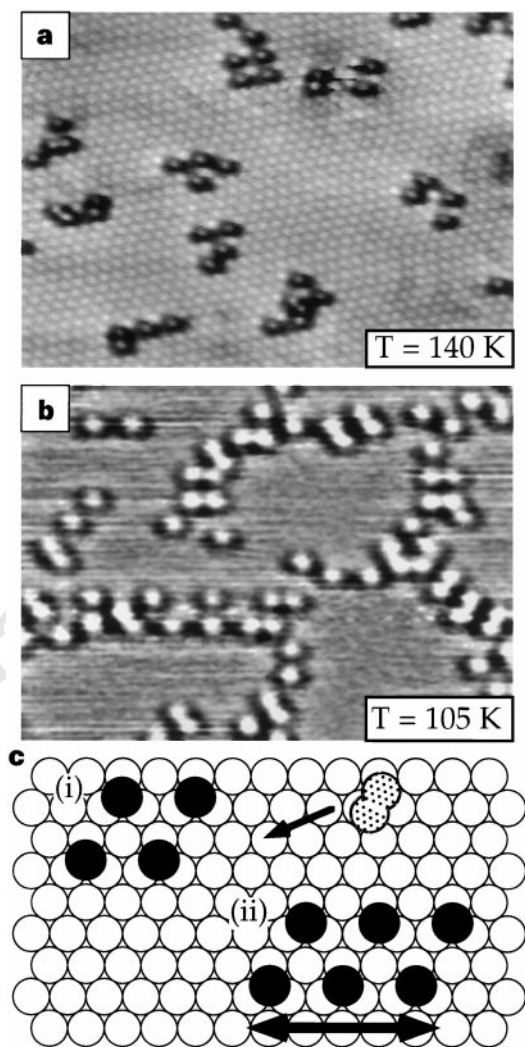


Figure 2 Atomic-resolution images of oxygen islands. The images illustrate the enhanced precursor dissociation probability at adsorbed oxygen atoms and the resulting quasi-one-dimensional growth for (a) $T = 140$ K (3 L dose, image $100 \times 60 \text{ \AA}^2$), (b) $T = 105$ K with chain-like islands corresponding to strings of oxygen pairs (1 L dose, image $70 \times 40 \text{ \AA}^2$). c, Model illustrating the directional growth of O_{ad} (black spheres) chains on Pt(111): (i) an oxygen precursor molecule (shaded) approaches an atom island; (ii) on its dissociation a chain segment evolves.

essentially induced by local modifications of the electronic properties of a surface near chemisorbed particles and hence, we believe, are of general importance for chemical processes at surfaces whenever the diffusion lengths of adsorbing species reach critical values compared with their lateral distribution. □

Received 12 May; accepted 7 October 1997.

1. Boudart, M. & Djega-Mariadassow, G. *Kinetics of Heterogeneous Catalytic Reactions* (Princeton Univ. Press, 1984).
2. Gland, J. L. Molecular and atomic adsorption of oxygen on the Pt(111) and Pt(S)-12(111) × (111) surfaces. *Surf. Sci.* **93**, 487–514 (1980).
3. Gland, J. L., Sexton, B. A. & Fisher, G. B. Oxygen interaction with the Pt(111) surface. *Surf. Sci.* **95**, 587–602 (1980).
4. Campbell, C. T., Ertl, G., Kuipers, H. & Segner, J. A molecular beam study of the adsorption and desorption of oxygen from a Pt(111) surface. *Surf. Sci.* **107**, 220–236 (1981).
5. Steininger, H., Lehwald, S. & Ibach, H. Adsorption of oxygen on Pt(111). *Surf. Sci.* **123**, 1–17 (1982).
6. Avery, N. R. An EELS and TDS study of molecular oxygen desorption and decomposition on Pt(111). *Chem. Phys. Lett.* **96**, 371–373 (1983).
7. Luntz, A. C., Williams, M. D. & Bethune, D. S. The sticking of O₂ on a Pt(111) surface. *J. Chem. Phys.* **89**, 4381–4395 (1988).
8. Williams, M. D., Bethune, D. S. & Luntz, A. C. Coexistence of precursor and direct dynamics: The sticking of O₂ on a Pt(111) surface. *J. Chem. Phys.* **88**, 2843–2845 (1988).
9. Winkler, A., Guo, X., Siddiqui, H. R., Hagans, P. L. & Yates, J. T. Kinetics and energetics of oxygen adsorption on Pt(111) and Pt(112)—a comparison of flat and stepped surfaces. *Surf. Sci.* **201**, 419–443 (1988).
10. Luntz, A. C., Grimblot, J. & Fowler, D. E. Sequential precursors in dissociative chemisorption: O₂ on Pt(111). *Phys. Rev. B* **39**, 12903–12906 (1989).
11. Wurth, W. *et al.* Bonding, structure, and magnetism of physisorbed and chemisorbed O₂ on Pt(111). *Phys. Rev. Lett.* **65**, 2426–2429 (1990).
12. Rettner, C. T. & Mullins, C. B. Dynamics of the chemisorption of O₂ on Pt(111): Dissociation via direct population of a molecularly chemisorbed precursor at high incidence kinetic energy. *J. Chem. Phys.* **94**, 1626–1635 (1991).
13. Puglia, C. *et al.* Physisorbed, chemisorbed and dissociated O₂ on Pt(111) studied with different core level spectroscopy methods. *Surf. Sci.* **342**, 119–133 (1995).
14. Wintterlin, J., Schuster, R. & Ertl, G. Existence of a “hot” atom mechanism for the dissociation of O₂ on Pt(111). *Phys. Rev. Lett.* **77**, 123–126 (1996).
15. Artsykhovich, A. N., Ukrainsev, V. A. & Harrison, I. Low temperature sticking and desorption dynamics of oxygen on Pt(111). *Surf. Sci.* **347**, 303–318 (1996).
16. Stipe, B. C. *et al.* Single-molecule dissociation by tunneling electrons. *Phys. Rev. Lett.* **78**, 4410–4413 (1997).
17. Witten, T. A. & Sander, L. M. Diffusion-limited aggregation, a kinetic critical phenomenon. *Phys. Rev. Lett.* **47**, 1400–1403 (1981).

Acknowledgements. The work of T.Z. was supported by the Deutscher Akademischer Austauschdienst (DAAD). Discussions with B.S. Stipe and A. C. Lutz are gratefully acknowledged.

Correspondence should be addressed to J.V.B. (e-mail: johannes.barth@ipe.dp.epfl.ch).

Effect of seawater carbonate concentration on foraminiferal carbon and oxygen isotopes

Howard J. Spero*, Jelle Bijma†, David W. Lea‡ & Bryan E. Bemis*

* Department of Geology, University of California Davis, Davis, California 95616, USA

† Alfred Wegener Institute for Polar and Marine Research, PO Box 120161, D-27515 Bremerhaven, Germany

‡ Department of Geological Sciences and the Marine Science Institute, University of California Santa Barbara, Santa Barbara, California 93106, USA

Stable oxygen and carbon isotope measurements on biogenic calcite and aragonite have become standard tools for reconstructing past oceanographic and climatic change. In aquatic organisms, ¹⁸O/¹⁶O ratios in the shell carbonate are a function of the ratio in the sea water and the calcification temperature¹. In contrast, ¹³C/¹²C ratios are controlled by the ratio of dissolved inorganic carbon in sea water and physiological processes such as respiration and symbiont photosynthesis². These geochemical proxies have been used with analyses of foraminifera shells to reconstruct global ice volumes³, surface and deep ocean temperatures^{4,5}, ocean circulation changes⁶ and glacial–interglacial exchange between the terrestrial and oceanic carbon pools⁷. Here, we report experimental measurements on living symbiotic

and non-symbiotic plankton foraminifera (*Orbulina universa* and *Globigerina bulloides* respectively) showing that the ¹³C/¹²C and ¹⁸O/¹⁶O ratios of the calcite shells decrease with increasing seawater [CO₃^{2−}]. Because glacial-period oceans had higher pH and [CO₃^{2−}] than today⁸, these new relationships confound the standard interpretation of glacial foraminiferal stable-isotope data. In particular, the hypothesis that the glacial–interglacial shift in the ¹³C/¹²C ratio was due to a transfer of terrestrial carbon into the ocean⁷ can be explained alternatively by an increase in ocean alkalinity²⁵. A carbonate-concentration effect could also help explain some of the extreme stable-isotope variations during the Proterozoic and Phanerozoic aeons⁹.

Data from isotopic calibration studies suggest that a previously unidentified parameter affects carbon and oxygen isotope ratios in foraminiferal shells¹⁰. Generally, observations of disequilibrium precipitation have been attributed to respiration or symbiont photosynthesis. These physiological processes contribute ¹³C-depleted CO₂ to the calcifying microenvironment (respiration) or enrich the environment in ¹³C via preferential removal of ¹²CO₂ (photosynthesis)². In contrast, shell ¹⁸O/¹⁶O ratios should be insensitive to physiology. Yet data demonstrate¹¹ that shell ^δ¹⁸O is often out of equilibrium with respect to temperature and ^δ¹⁸O_{water} (see Methods). Laboratory and field data show that the ^δ¹⁸O and ^δ¹³C of CaCO₃ can covary^{11,12}, suggesting that a common mechanism affects the isotope ratios of both elements.

Our experimental results demonstrate that at constant alkalinity, but varying ΣCO₂, the ^δ¹³C and ^δ¹⁸O of *O. universa* shells decrease by about 3.9‰ and 1.5‰ respectively as [CO₃^{2−}] increases from 100 to 642 μmol kg^{−1} (see Methods and Fig. 1a, b). Shell ^δ¹³C and ^δ¹⁸O values remain constant when seawater [CO₃^{2−}] is further reduced from 100 to 41 μmol kg^{−1}. Comparison of foraminifera grown under photosynthetic maximum (*P*_{max}) light levels¹³ (denoted HL) with specimens maintained in the dark (D) (to compare maximum with minimum symbiont photosynthetic activity), shows that *P*_{max} shells are consistently enriched in ¹³C by 1.1‰ and depleted in ¹⁸O by 0.3‰ relative to dark specimens. Although this effect has been documented previously¹⁴, the constant offset, and the identical regression slopes (−0.0063 ± 0.0007 (HL) (95% confidence interval) and −0.0061 ± 0.0013‰/(μmol kg^{−1}) (D) for ^δ¹³C/[CO₃^{2−}], and −0.0022 ± 0.0006 (HL) and −0.0019 ± 0.0002‰/(μmol kg^{−1}) (D) for ^δ¹⁸O/[CO₃^{2−}]) demonstrate that symbiont photosynthesis is not responsible for the observed relationships between isotopes and [CO₃^{2−}].

In a second experiment, we manipulated [CO₃^{2−}] between 75 and 774 μmol kg^{−1} by varying alkalinity and maintaining ΣCO₂ at 2,032 ± 15 μmol kg^{−1} (Fig. 1c and d). Again, ^δ¹³C and ^δ¹⁸O in *O. universa* shells decreased as [CO₃^{2−}] increased, with regression slopes of −0.0055 ± 0.0015 (HL) and −0.0060 ± 0.0015‰/(μmol kg^{−1}) (D) for carbon and −0.0015 ± 0.0008 (HL) and −0.0022 ± 0.0004‰/(μmol kg^{−1}) (D) for oxygen. For a companion experiment, several *O. universa* were removed from sea water of high [CO₃^{2−}] (458 μmol kg^{−1}) after sphere formation but before gametogenesis to test whether the carbonate ion effect was due to inorganic precipitation on the surface of empty, post-gametogenic shells. The ^δ¹³C and ^δ¹⁸O values of both pre- and post-gametogenic shells were identical, indicating that the relationship is a function of calcification during normal shell growth. Furthermore, because the seawater carbonate chemistry in our experiments was modified two ways yet produced indistinguishable regression slopes, the relationship between isotopic ratios and [CO₃^{2−}] is real and not an artefact of our experiments.

In a final experiment, we cultured a nonsymbiotic species, *G. bulloides*, across a [CO₃^{2−}] range of 103–436 μmol kg^{−1} at constant ΣCO₂. Because *G. bulloides* displays a large chamber-to-chamber ontogenetic effect for both ^δ¹³C and ^δ¹⁸O¹¹, we amputated those chambers secreted during the experiments and pooled chambers from discrete positions in the shell whorl for isotope measurement

(Fig. 2). Regression analysis of data from the 12th and 13th chambers yields slopes of -0.012 to -0.014 for $\delta^{13}\text{C}/[\text{CO}_3^{2-}]$ and -0.004 to -0.005 for $\delta^{18}\text{O}/[\text{CO}_3^{2-}]$, demonstrating that ontogeny has little affect on this relationship. Interestingly, the *G. bulloides* regression slopes are twice those of *O. universa*, suggesting that the slope of the carbonate ion effect is species-specific.

We recognize that these experiments extend beyond the present seawater $[\text{CO}_3^{2-}]$ range as well as that proposed for the last glacial maximum⁸. Statistical analysis of data from the constant-alkalinity experiment on *O. universa*, which encompasses the proposed Quaternary $[\text{CO}_3^{2-}]$ range ($100\text{--}400\ \mu\text{mol kg}^{-1}$), yields slopes that are indistinguishable (within 95% confidence intervals) from the full data sets with exception of the dark $\delta^{13}\text{C}/[\text{CO}_3^{2-}]$ slope of $-0.0033\text{‰}/(\mu\text{mol kg}^{-1})$. Because we have higher confidence in the stable isotope/ $[\text{CO}_3^{2-}]$ slopes calculated from the entire experimental range, we use those values for discussion in the following sections.

Evidence for a similar isotope effect exists in corals and other invertebrate groups that display a $\delta^{18}\text{O}:\delta^{13}\text{C}$ covariance¹⁵. For instance, in the nonphotosynthetic coral, *Tubastrea*, isotope values covary with a positive $\delta^{18}\text{O}/\delta^{13}\text{C}$ slope of 0.29. The foraminiferal $\delta^{18}\text{O}/\delta^{13}\text{C}$ relationships obtained here are virtually identical, with slopes ranging between 0.29 and 0.33 (Fig. 3). Other ahermotypic corals, calcareous algae and invertebrates such as cidaroid urchins show oxygen and carbon isotope covariance^{15,16}, although the slopes can differ from the experimental range. In symbiont-

bearing organisms such as the coral *Pavona* or foraminifer *Globigerinoides sacculifer*, the slope can change sign because of the additional ^{13}C -enriching effect of symbiont photosynthesis^{15,17}. Aside from organisms that display a large symbiont effect, the similarities in the slope of the relationship among different protozoans, invertebrates and calcifying algae suggests that related mechanisms are responsible for the isotope covariance.

Comparison of our experimental data with results for inorganic precipitates¹⁸ shows remarkable similarity in $\delta^{18}\text{O}$ variations with $[\text{CO}_3^{2-}]$ (Fig. 4). McCrea demonstrated that the $^{18}\text{O}/^{16}\text{O}$ ratio of rapidly precipitated CaCO_3 decreases with increasing percentage of CO_3^{2-} in solution and suggested that the calcification rate plays a role. The similarity between biological and inorganic precipitate studies implicates a kinetic mechanism which may affect all calcifying organisms and inorganic precipitates. Although we cannot estimate the calcification rates of *O. universa* in our experiments, shell mass measurements on high light and dark groups show that specimens grown in high $[\text{CO}_3^{2-}]$ (over $600\ \mu\text{mol kg}^{-1}$) are 37% heavier than those in ambient sea water. Because the *O. universa* lifespan was similar for the different treatments, heavier shells probably indicate higher calcification rates. Chamber mass data for *G. bulloides* are inconclusive.

Although we cannot offer a definitive mechanism for the isotope: $[\text{CO}_3^{2-}]$ relationships, mass balance calculations show that the results cannot be explained by a simple redistribution of isotopes between the dissolved carbon species¹⁹. McConnaughey¹²

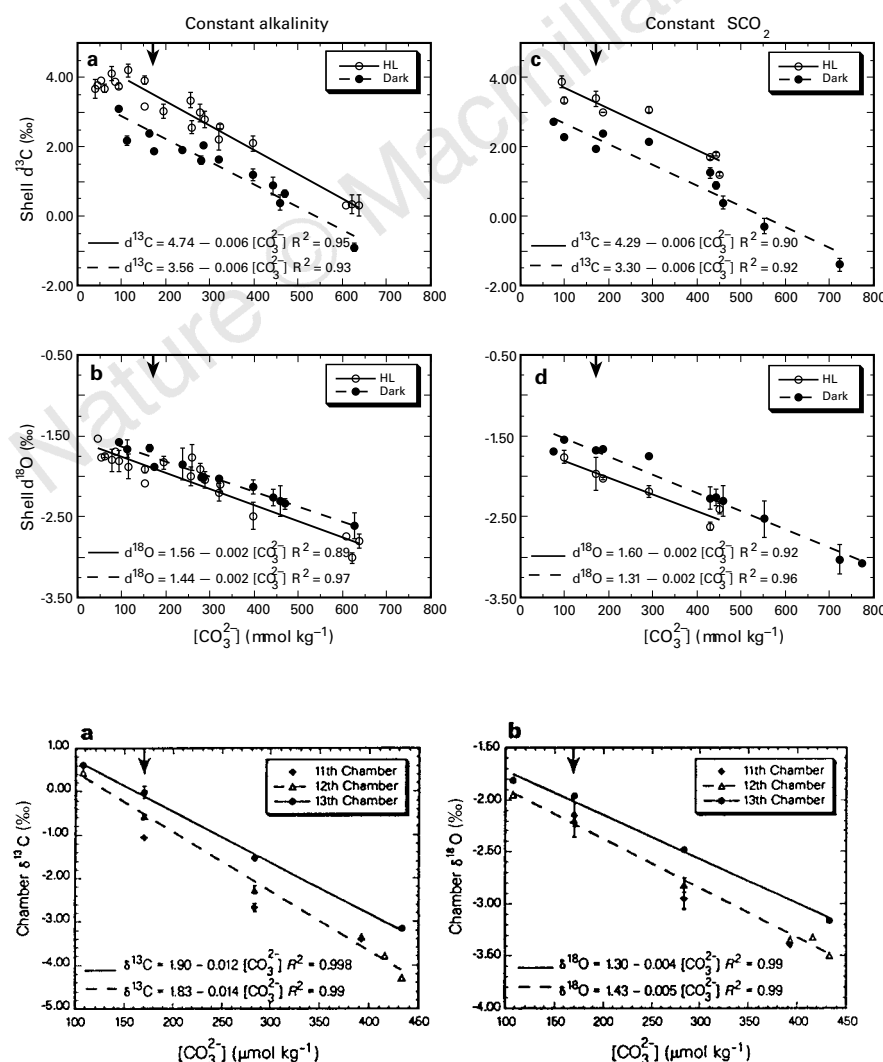


Figure 1 Effect of $[\text{CO}_3^{2-}]$ on the $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ values of *Orbulina universa* shell calcite under conditions of constant alkalinity (a and b) and constant ΣCO_2 (c and d) conditions. Specimens grown under high light and in the dark are shown by open and closed circles respectively. Arrowheads identify ambient $[\text{CO}_3^{2-}]$ in the Southern California Bight. Data are group mean values ± 1 s.d.; most groups are composed of 3–10 individual shells. Lines are linear regressions fitted to the data.

Figure 2 Effect of $[\text{CO}_3^{2-}]$ on the $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ values of *Globigerina bulloides* chamber calcite under constant ΣCO_2 conditions. Data are from pooled, amputated chambers from specific positions in the shell whorl. Linear regression analyses are plotted for data from chambers 12 and 13, the last two chambers in the shell.

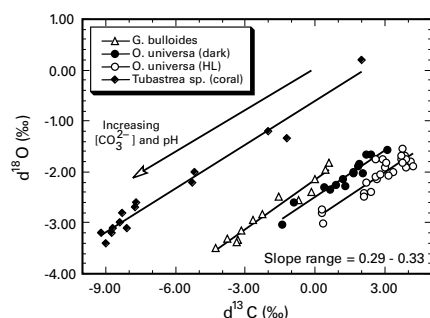


Figure 3 Comparison of foraminiferal $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ data from all experiments with data from the nonsymbiotic coral, *Tubastrea* spp. (ref. 15). The similarity of the positive covariance between coral (aragonite) and foraminifera (calcite) stable isotopes suggests that a similar mechanism is responsible for the observed variations. In this plot, $[\text{CO}_3^{2-}]$ and pH increase towards the origin.

has explained the strong $\delta^{13}\text{C}$: $\delta^{18}\text{O}$ covariance in the coral *Tubastrea* by a combination of 'kinetic' and 'metabolic' isotope effects. The kinetic effect is attributed to slower CO_2 hydration and hydroxylation reactions by molecules enriched in the heavier isotopes ^{13}C and ^{18}O . McConnaughey further argues that CO_2 hydration and hydroxylation reactions may show different kinetic isotope effects and that the balance between these two reactions changes with pH. At high pH and $[\text{CO}_3^{2-}]$, CO_3^{2-} favours the lighter isotope in the exchange of oxygen atoms between CO_2 and CO_3^{2-} . A similar mechanism was proposed by McCrea¹⁸ and has been supported by more recent work²⁰.

In contrast, metabolic effects can be attributed to the incorporation of ^{13}C -depleted respired CO_2 . For aquatic organisms, such respiratory effects may be small and difficult to evaluate against the larger kinetic effect²¹. In some benthic foraminifera, an internal inorganic carbon pool can introduce added complexity when one attempts to resolve the carbon sources controlling shell $\delta^{13}\text{C}$ ²². Although a carbon pool is absent in *O. universa* (J.B., unpublished data), metabolic carbon does contribute 8–15% to *G. bulloides* shell carbon¹¹. Such differences in carbon sources might explain the species-specific $\delta^{13}\text{C}/[\text{CO}_3^{2-}]$ slopes observed here but would not be expected to influence $\delta^{18}\text{O}$. Current experiments should help to constrain the active mechanism(s).

Because the CO_2 partial pressure (p_{CO_2}) of the sea surface and the atmosphere must be in approximate equilibrium, the glacial drop in atmospheric carbon dioxide recorded in ice cores^{23,24} must have been accompanied by an increase in surface water $[\text{CO}_3^{2-}]$. Therefore shells calcifying during cold periods record isotopic signatures that cannot be interpreted using present-day relationships. Our calculations indicate that to achieve equilibrium with a glacial atmospheric p_{CO_2} of 200 parts per million by volume (p.p.m.V), surface water $[\text{CO}_3^{2-}]$ must have increased by a minimum of $40 \mu\text{mol kg}^{-1}$, with considerably larger changes possible depending on sea surface temperature (SST) and the actual mode of p_{CO_2} change²⁵. Combining this conservative estimate with the experimental relationships presented here suggests that the glacial rise in $[\text{CO}_3^{2-}]$ can account for at least a 0.25–0.5‰ drop in shell $\delta^{13}\text{C}$, with the larger change characterizing *G. bulloides*. Large glacial-to-interglacial (G–I) shifts, 0.7‰ or greater, observed in $\delta^{13}\text{C}$ records from *G. bulloides*²⁶ are consistent with the greater response of this species to $[\text{CO}_3^{2-}]$. The smaller glacial shifts recorded in tropical records of *G. sacculifer* (0.3‰)²⁷ suggest a $[\text{CO}_3^{2-}]$ response for this species that is similar to *O. universa*, a hypothesis that can be tested experimentally. If some or all of the negative carbon isotope shift that is common to all species can be attributed to an increase in $[\text{CO}_3^{2-}]$, this reduces the actual G–I mean shift in $\delta^{13}\text{C}$ for the ocean. This would enlarge the difference between pollen-based and ocean- $\delta^{13}\text{C}$ -based estimates of the amount of carbon transferred from the terrestrial biosphere to

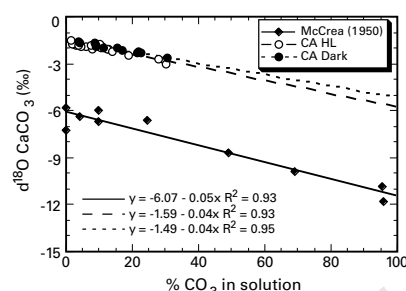


Figure 4 Comparison of high light and dark *O. universa* $\delta^{18}\text{O}$ data from the constant alkalinity (CA) experiment with inorganic precipitate results from McCrea¹⁸. Oxygen isotope data are plotted against % CO_3^{2-} in solution, where $\% \text{CO}_3^{2-} = [(\text{CO}_3^{2-})/((\text{CO}_3^{2-}) + [\text{HCO}_3^-]) \times 100$. Note that the slopes of our experimental data are indistinguishable from that obtained by McCrea.

the ocean during glacial episodes^{7,28}.

The influence of $[\text{CO}_3^{2-}]$ also confounds the interpretation of $\delta^{13}\text{C}$ differences between planktic and benthic foraminifera as a measure of the efficiency of the biological pump²⁶. Because a 0.1‰ increase in $\Delta\delta^{13}\text{C}$ translates to an approximate 10 p.p.m.V drop in p_{CO_2} , a 0.3‰ correction on planktics, which is the minimum calculated, can explain an additional glacial-interglacial p_{CO_2} difference of 30 p.p.m.V, implying a stronger glacial biological pump than previously indicated²⁷. However, this calculation depends on the magnitude of the $[\text{CO}_3^{2-}]$ influence on planktic species for which experimental data are not available, and it must also take into account a potential $[\text{CO}_3^{2-}]$ influence on benthic shell $\delta^{13}\text{C}$ which has not yet been demonstrated. Correcting for the more subtle influence of $[\text{CO}_3^{2-}]$ on shell $\delta^{18}\text{O}$ lowers glacial tropical SST estimates by up to 1 °C, bringing oxygen isotope palaeotemperatures closer to those calculated from other marine SST proxies (such as coral Sr/Ca ratios) and terrestrial indicators (such as snowline and ice-core $\delta^{18}\text{O}$) that suggest more intense tropical cooling at the last glacial maximum^{29,30}.

The late Proterozoic and early Phanerozoic aeons must have had very different ocean carbonate chemistries from the present in order to maintain high p_{CO_2} and accumulate extensive carbonate platforms^{31,32}. Higher ΣCO_2 , alkalinity and/or $[\text{CO}_3^{2-}]$ could satisfy both requirements. If a carbonate ion effect were operating, higher $[\text{CO}_3^{2-}]$ would result in CaCO_3 with lower $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ values than predicted from current ocean chemistry. Large negative $\delta^{13}\text{C}$ excursions have been recorded throughout the late Proterozoic, especially near the Vendian/Cambrian boundary⁹. In the early Palaeozoic, very low $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ values have been recovered from articulate brachiopods³³. Could changes in ocean carbonate chemistry help to explain some of these observations? □

Methods

Collection and culturing. *O. universa* and *G. bulloides* were collected by scuba divers from surface waters of the San Pedro Basin, Southern California Bight, USA, and were maintained in laboratory culture at $22 \pm 0.2^\circ\text{C}$ at the Wrigley Institute of Environmental Science on Santa Catalina Island. Specimens were grown in 0.8- μm -filtered sea water in sealed, air-tight 125 mL acid-cleaned Pyrex jars and fed a one-day-old *Artemia* nauplius every third day (total time in culture ~6–8 days). For *O. universa*, specimens were maintained either in the dark or under P_{max} light levels ($400\text{--}700 \mu\text{Ein m}^{-2} \text{s}^{-1}$; 1 einstein = 1 mole photons)¹³ to quantify the effect of symbiont photosynthesis on the relationship between stable isotopes and $[\text{CO}_3^{2-}]$. Following gametogenesis, the empty foraminiferal shells were archived for analysis. *G. bulloides* chambers secreted during the experiments were amputated from the shell whorl and pooled according to chamber position¹¹. In *O. universa*, a single spherical chamber was secreted in the laboratory which accounted for 90–95% of total shell calcite.

Seawater chemistry manipulation. We modified seawater carbonate

chemistry by either: (1) raising total alkalinity (alk_t) (where alkalinity = $[\text{HCO}_3^-] + 2[\text{CO}_3^{2-}] + [\text{HBO}_3^-]$) to constant levels of $2,842 \pm 80 \mu\text{eq kg}^{-1}$ (equivalent = mole of negative or positive charge) ($n = 29$ groups) (ambient $\text{alk}_t = 2,241 \pm 19 \mu\text{eq kg}^{-1}$; $n = 64$ groups) and letting ΣCO_2 ($\Sigma\text{CO}_2 = [\text{CO}_2] + [\text{HCO}_3^-] + [\text{CO}_3^{2-}]$) and pH vary (pH ranges between 7.38 and 8.83); or (2) keeping ΣCO_2 constant at $2,032 \pm 15 \mu\text{mol kg}^{-1}$ ($n = 15$ groups) (ambient $\Sigma\text{CO}_2 = 2,010 \pm 18 \mu\text{mol kg}^{-1}$; $n = 64$ groups) and letting pH and alkalinity vary (pH ranges from 7.87 to 8.97). For comparison with our experiments, pH and $[\text{CO}_3^{2-}]$ in Southern California Bight surface waters varied between 8.11–8.19 and $153\text{--}184 \mu\text{mol kg}^{-1}$ during our field campaign.

Seawater ΣCO_2 and alk_t were modified by the addition of Na_2CO_3 and/or titration with HCl or NaOH to bring alk_t to desired levels. Seawater pH was determined by potentiometry whereas alk_t and ΣCO_2 were determined by titration and equilibrium calculations respectively. Coulometric determinations of several seawater samples confirm the accuracy of the calculated ΣCO_2 values (A. Sanyal, personal communication). Culture water samples collected at the start and end of each experiment showed that alkalinity and ΣCO_2 remained constant throughout each experiment.

Stable isotope analyses. Individual *O. universa* shells and *G. bulloides* chambers were roasted at 375°C in *vacuo* and analysed with a Fisons Optima isotope ratio mass spectrometer using an Isocarb common acid bath autocarbonate system at 90°C . Here $\delta^{18}\text{O} = [({}^{18}\text{O}/{}^{16}\text{O}_{\text{sample}}/{}^{18}\text{O}/{}^{16}\text{O}_{\text{std}}) - 1]$ and $\delta^{13}\text{C} = [({}^{13}\text{C}/{}^{12}\text{C}_{\text{sample}}/{}^{13}\text{C}/{}^{12}\text{C}_{\text{std}}) - 1]$. All $\delta^{13}\text{C}_{\Sigma\text{CO}_2}$ and CaCO_3 isotope values are relative to the V-PDB standard. Water samples collected at the start and end of each experiment show that $\delta^{18}\text{O}_{\text{water}}$ was constant at $-0.23 \pm 0.05\text{‰}$ (relative to the V-SMOW standard) ($n = 100$) whereas initial and final water $\delta^{13}\text{C}_{\Sigma\text{CO}_2}$ differed on average by $0.16 \pm 0.10\text{‰}$ ($n = 30$). If foraminiferal $\delta^{13}\text{C}$ data have been corrected to a constant $\delta^{13}\text{C}_{\Sigma\text{CO}_2} = 2.00\text{‰}$ (ambient $\delta^{13}\text{C}_{\Sigma\text{CO}_2}$ was $1.90 \pm 0.08\text{‰}$, $n = 18$) to account for $\delta^{13}\text{C}_{\Sigma\text{CO}_2}$ differences between treatments due to the addition of Na_2CO_3 .

Received 6 January; accepted 7 October 1997.

- Epstein, S., Buchsbaum, R., Lowenstam, H. A. & Urey, H. C. Revised carbonate-water isotopic temperature scale. *Geol. Soc. Am. Bull.* **64**, 1315–1325 (1953).
- Spero, H. J., Lerche, I. & Williams, D. E. Opening the carbon isotope “vital effect” black box, 2: Quantitative model for interpreting foraminifera carbon isotope data. *Paleoceanography* **6**, 639–655 (1991).
- Shackleton, N. J. & Opdyke, N. D. Oxygen isotope and palaeomagnetic stratigraphy of equatorial Pacific core V28-238: Oxygen isotope temperatures and ice volumes on a 10^5 year and 10^6 year scale. *Quat. Res.* **3**, 39–55 (1973).
- Broecker, W. S. Oxygen isotope constraints on surface ocean temperatures. *Quat. Res.* **26**, 121–134 (1986).
- Labeyrie, L. D., Duplessy, J.-C. & Blanc, P. L. Variations in mode of formation and temperature of oceanic deep waters over the past 125,000 years. *Nature* **327**, 477–482 (1987).
- Duplessy, J. C. et al. Deepwater source variations during the last climatic cycle and their impact on the global deepwater circulation. *Paleoceanography* **3**, 343–360 (1988).
- Shackleton, N. J. Carbon-13 in Uvigerina: tropical rainforest history and the equatorial Pacific carbonate dissolution cycles. In *The Fate of Fossil Fuel CO₂ in the Oceans* (eds Andersen, N. R. & Malahoff, A.) 401–427 (Plenum, New York, 1977).
- Sanyal, A., Hemming, N. G., Hanson, G. N. & Broecker, W. S. Evidence for a higher pH in the glacial ocean from boron isotopes in foraminifera. *Nature* **373**, 234–236 (1995).
- Kaufman, A. J., Jacobsen, S. B. & Knoll, A. H. The Vendian record of Sr and C isotopic variations in seawater: Implications for tectonics and paleoclimate. *Earth Planet. Sci. Lett.* **120**, 409–430 (1993).
- Fairbanks, R. G., Sverdrup, M., Free, R., Wiebe, P. H. & Be, A. W. H. Vertical distribution and isotopic fractionation of living planktonic foraminifera from the Panama Basin. *Nature* **298**, 841–844 (1982).
- Spero, H. J. & Lea, D. W. Experimental determination of stable isotope variability in *Globigerina bulloides*: Implications for paleoceanographic reconstruction. *Mar. Micropaleontol.* **28**, 231–246 (1996).
- McConnaughey, T. ^{13}C and ^{18}O isotopic disequilibrium in biological carbonates: II. In vitro simulation of kinetic isotope effects. *Geochim. Cosmochim. Acta* **53**, 163–171 (1989).
- Spero, H. J. & Parker, S. L. Photosynthesis in the symbiotic planktonic foraminifer *Orbulina universa*, and its potential contribution to oceanic primary productivity. *J. Foram. Res.* **15**, 273–281 (1985).
- Spero, H. J. Do planktic foraminifera accurately record shifts in the carbon isotopic composition of sea water ΣCO_2 ? *Mar. Micropaleontol.* **19**, 275–285 (1992).
- McConnaughey, T. ^{13}C and ^{18}O isotopic disequilibrium in biological carbonates: I. Patterns. *Geochim. Cosmochim. Acta* **53**, 151–162 (1989).
- Smith, J. E., Risk, M. J., Schwarcz, H. P. & McConnaughey, T. A. Rapid climate change in the North Atlantic during the Younger Dryas recorded by deep-sea corals. *Nature* **386**, 818–820 (1997).
- Spero, H. J. & Lea, D. W. Intraspecific stable isotope variability in the planktic foraminifera *Globigerinoides sacculifer*: Results from laboratory experiments. *Mar. Micropaleontol.* **22**, 221–234 (1993).
- McCrea, J. M. On the isotopic chemistry of carbonates and a paleotemperature scale. *J. Chem. Phys.* **18**, 849–857 (1950).
- Zhang, J., Quay, P. D. & Wilbur, D. O. Carbon isotope fractionation during gas–water exchange and dissolution of CO_2 . *Geochim. Cosmochim. Acta* **59**, 107–114 (1995).
- Udowski, E. & Hoefs, J. Oxygen isotope exchange between carbonic acid, bicarbonate, carbonate, and water: A re-examination of the data of McCrea (1950) and an expression for the overall partitioning of oxygen isotopes between the carbonate species and water. *Geochim. Cosmochim. Acta* **57**, 3815–3818 (1993).

- McConnaughey, T. A., Burdett, J., Whelan, J. F. & Paull, C. K. Carbon isotopes in biological carbonates: Respiration and photosynthesis. *Geochim. Cosmochim. Acta* **61**, 611–622 (1997).
- ter Kuile, B., Erez, J. & Padan, E. Mechanisms for the uptake of inorganic carbon by two species of symbiont-bearing foraminifera. *Mar. Biol.* **103**, 241–251 (1991).
- Barnola, J. M., Raynaud, D., Korotkevich, Y. S. & Lorius, C. Vostock ice core provides 160,000-year record of atmospheric CO_2 . *Nature* **329**, 408–413 (1987).
- Broecker, W. S. & Peng, T.-H. What caused the glacial to interglacial CO_2 change? in *The Global Carbon Cycle* (ed. Heimann, M.) 95–115 (Springer, Berlin, 1993).
- Lea, D. W., Spero, H. J., Bijma, J. & Archer, D. Implications of a carbonate ion effect on shell carbon and oxygen isotopes for glacial ocean conditions. *EOS* **77** (46) Fall Meeting Suppl. F334 (1996).
- Shackleton, N. J., Hall, M. A., Line, J. & Shuxi, C. Carbon isotope data in core V19-30 confirm reduced carbon dioxide concentration in the ice age atmosphere. *Nature* **306**, 319–322 (1983).
- Shackleton, N. J., Le, J., Mix, A. & Hall, M. A. Carbon isotope records from Pacific surface waters and atmospheric carbon dioxide. *Quat. Sci. Rev.* **11**, 387–400 (1992).
- Crowley, T. J. Ice age terrestrial carbon changes revisited. *Global Biogeochem. Cycles* **9**, 377–389 (1995).
- Guilderson, T. P., Fairbanks, R. G. & Rubenstein, J. L. Tropical temperature variations since 20,000 years ago: Modulating interhemispheric climate change. *Science* **263**, 663–665 (1994).
- Thompson, L. G. et al. Late glacial stage and Holocene tropical ice core records from Huascaran, Peru. *Science* **269**, 46–50 (1995).
- Berner, R. A. Geocarb II: A revised model of atmospheric CO_2 over Phanerozoic time. *Am. J. Sci.* **294**, 56–91 (1994).
- Grotzinger, J. P. & Kasting, J. F. New constraints on Precambrian ocean composition. *J. Geol.* **101**, 235–243 (1993).
- Wadleigh, M. A. & Veizer, J. $^{18}\text{O}/^{16}\text{O}$ and $^{13}\text{C}/^{12}\text{C}$ in Lower Paleozoic articulate brachiopods: Implications for the isotopic composition of seawater. *Geochim. Cosmochim. Acta* **56**, 431–443 (1992).

Acknowledgements. We thank the staff of the Wrigley Institute of Environmental Science and E. Kincaid, C. Hamilton, J. Dailey, E. Komsky, T. Mashiotto, M. Uhle, A. Sanyal, D. Chan, E. Mochon and M. Cramer for their help in the field. Thanks also to A. Russell and D. Sumner for comments on the manuscript. This research was supported by the US National Science Foundation (H.J.S. and D.W.L.) and by SFB 261 and the Program for the Advancement of Special Research Projects at the Alfred Wegener Institute, Germany (J.B.).

Correspondence should be addressed to H.J.S. (e-mail: spero@geology.ucdavis.edu).

Localization of the gravity field and the signature of glacial rebound

Mark Simons* & Bradford H. Hager

Department of Earth Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

The negative free-air gravity anomaly centred on Hudson Bay, Canada, shows a remarkable correlation with the location of the Laurentide ice sheet, suggesting that this gravity anomaly is the result of incomplete post-glacial rebound^{1–3}. This region, however, is also underlain by higher-than-average mantle seismic velocities, suggesting that the gravity low might result instead from dynamic topography associated with convective downwellings^{4–7}. Here we analyse the global gravity field as a simultaneous function of geographic location and spectral content. We find that the Hudson Bay gravity low is unique, with anomalously high amplitude in the spectral band where the power from the Laurentide ice load is greatest² and the relaxation times predicted for viable models of viscous relaxation are longest⁸. We estimate that about half of the Hudson Bay gravity anomaly is the result of incomplete post-glacial rebound, and derive a mantle viscosity model that explains both this gravity signature and the characteristic uplift rates for the central Laurentide and Fennoscandian regions⁶. This model has a jump in viscosity at 670 km depth, comparable to that in dynamic models of the geoid highs over subducted slabs^{4,9}, but lacks a low-viscosity asthenosphere, consistent with a higher viscosity in the upper mantle beneath shields than in oceanic regions.

Delayed rebound is not the dominant process generating variations in the global gravity field (Fig. 1a); many anomalies of comparable or greater amplitude than those in deglaciated areas

* Present address: Seismological Laboratory, California Institute of Technology, Pasadena, California 91125, USA.

exist elsewhere. Indeed, there is a poor global correlation between the observed gravity field and that predicted by models of glacial isostasy^{10,11}. Previous attempts to relate the observed gravity field to the glacial rebound process have focused either on the local amplitude of the gravity field^{7,11} or on global correlations as a function of spherical harmonic degree^{10,11}. Our approach relies on a spatio-spectral localization method for spherical harmonic representations of global data sets^{12,13}. This method is similar to wavelet techniques in the cartesian domain and is well suited to global geophysical data. The method can be expressed as spatial windowing followed by spectral decomposition. We use a smooth axisymmetric window with a characteristic spatial width equal to twice the wavelength being considered. The window is translated over the globe providing sets of localized coefficients at all positions and wavelengths. Further details on this procedure, including the limitations imposed by using data with finite resolution, can be found elsewhere^{12,13}.

Maps of the spatially and spectrally localized r.m.s. amplitude of

the gravity field, $S_l(\Omega)$, are shown in Fig. 1b. At $l = 4$ there is a broad maximum over central Asia, Indonesia, Australia and Antarctica, and two minima over Africa and the central Pacific. Regions of subduction are frequently associated with S_l highs. At $l \leq 9$ this high is associated with mantle convection processes^{12,14} and at $l \geq 25$ with deep sea trenches. S_9 has a maximum over Hudson Bay. For $l \geq 9$, there are S_l maxima over the Andes, the Tibetan plateau and regions north of the plateau.

Spectrograms of the gravity field along two great-circle paths are shown in Fig. 1c–e. Hudson Bay is unique in that it is characterized by a gravity anomaly that is localized both in space and in length scale, with a peak ΔS_l value at $l \approx 9$, corresponding to length scales of $\sim 4,200$ km. In contrast with Hudson Bay, regions containing steep gradients in the gravity field, such as the Himalayas, are characterized by spatially localized spectral ridges. For the Himalayas, the spectral ridge extends down to $l \approx 10$, or to wavelengths as great as 4,000 km. The Sri Lankan low is part of a larger-length-scale low, extending from south of Australia to central Asia, that is

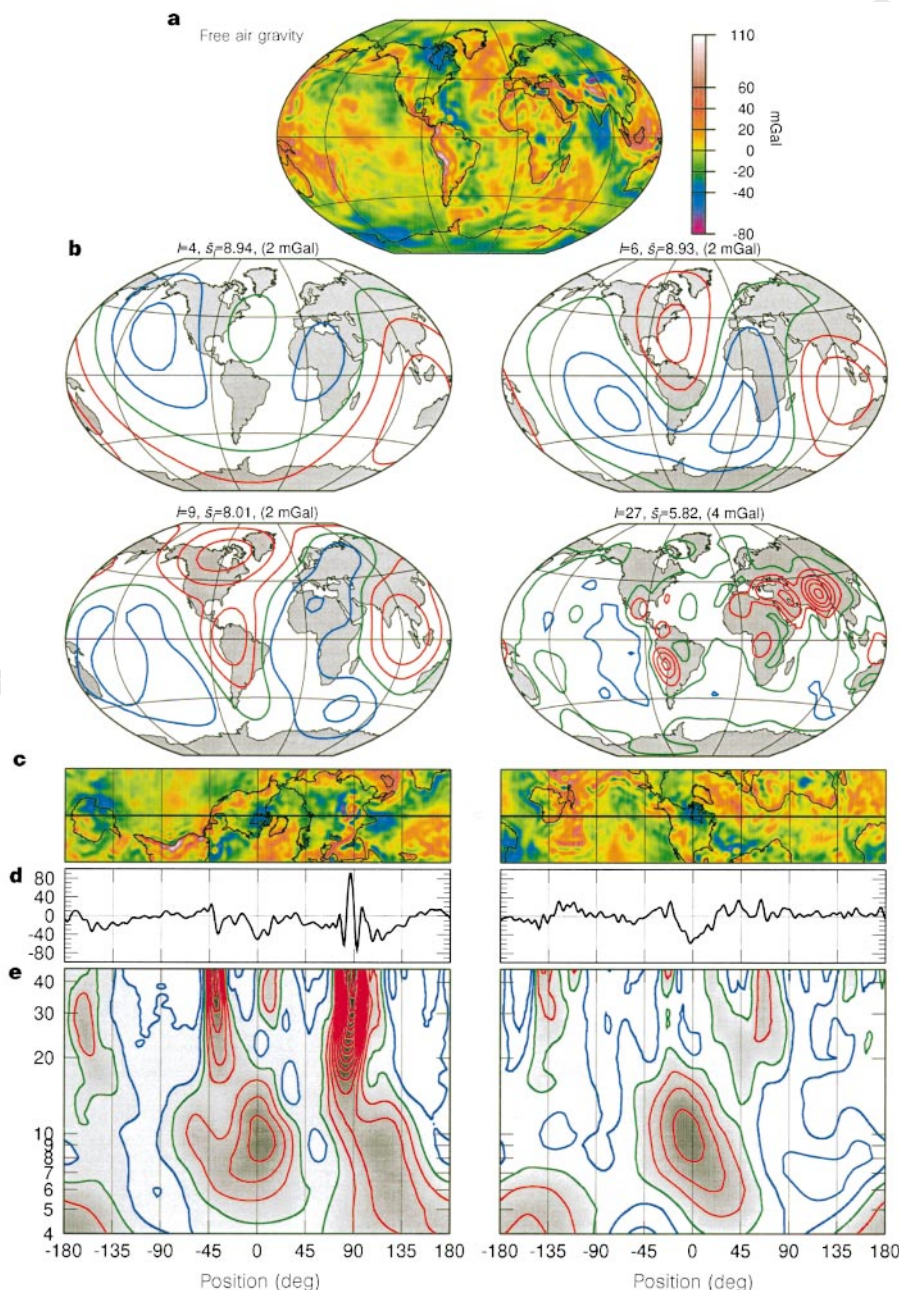


Figure 1 Spatio-spectral renditions of the gravity field.

a, Spatial rendition of the free air gravity field from spherical harmonic model JGM-2G (refs 24, 25). For this figure, the coefficients for the gravity field have been smoothly tapered from $l = 50$ –70. The analysis presented here has been conducted with both a free-air gravity and a geoid representation. Although we focus on the former representation, the conclusions do not depend on this choice. A Winkel Tripel projection centred at 45° W is used for all global maps shown here. **b**, Maps of $\Delta S_l(\Omega)$ for selected l , where $\Delta S_l(\Omega) = S_l(\Omega) - \bar{S}_l$ and $\bar{S}_l$ is the global average of $S_l(\Omega)$. $\bar{S}_l$ and the contour interval are indicated at the top of each diagram. $\Delta S_l > 0$, $\Delta S_l = 0$ and $\Delta S_l < 0$ are indicated by red, green and blue contours, respectively. Note that ΔS_l is a measure of the variation, but not the sign, of the original signal. **c**, Location maps for the two great-circle S_l spectrograms (chosen to be nearly orthogonal at Hudson Bay). Oblique Mercator projection for which the Equator is the great circle path indicated by the horizontal line. **d**, Gravity profiles along the great-circle paths. The localization uses the full two-dimensional field. **e**, ΔS_l as a function of l and position along the great circle. Same line convention as in **b**, with a 2 mGal contour interval²⁶.

interrupted by a short-length-scale positive anomaly over the Himalayas and Tibet. The short-length-scale anomaly is presumably related to crustal thickness variations, whereas the longer-length-scale negative anomaly is due to deeper convective processes. Unlike the Hudson Bay low, the Sri Lankan low has a spectral maximum extending down to the smallest l considered. Another spectral ridge, extending down to $l \approx 13$, is found traversing central America, again presumably related to the large variation in crustal thickness. Over the same region, S_l has a minimum at $l \approx 10$ and then increases again at lower l , due to the effects of the subducting slab at the Middle American subduction zone¹⁴. In contrast with the large S_l associated with deglaciated regions and subduction zones, we find no obvious signals at low degrees associated with hotspots. The South Pacific area is notable because it appears as a minimum in the estimates of S_l of the gravity field at all wavelengths.

Unlike for the gravity field, for seismic tomography models (for example, ref. 15) at depths between 100 and 300 km, estimates of S_l at $l = 6-11$ do not have a unique maximum associated with Hudson Bay (see Supplementary Information). Therefore, although there is probably a convective component to the total gravity anomaly in Hudson Bay, a significant portion of this anomaly is apparently related to glacial rebound. Unfortunately, it is difficult to separate the individual contributions to the gravity field from convection and incomplete glacial rebound. Furthermore, models of both of these processes are highly dependent on the rheological structure assumed^{4,9,14,16-19}, and the success of these models has been limited by their reliance on global variance reduction or spatial amplitude to measure their performance. By necessity, this reliance implies that one is also modelling gravity anomalies that are unrelated to the

process being considered. Here, we take an approach that is based on the local correlation between the observed gravity field, $N(\Omega)$, and the spatial distribution of a tectonic process, $T(\Omega)$, in this case the amplitude of deglaciation at a particular position. We calculate the local l -dependent transfer function, $F_l(\Omega)$, between $T(\Omega)$ and $N(\Omega)$ over the entire planet. A single global transfer function, $\bar{F}_l$, is derived by spatially averaging $F_l(\Omega)$ weighted by the local variance, $S_l^2(\Omega)$, of $T(\Omega)$, to emphasize regions that have experienced deglaciation. A predicted gravity field can then be created by convolving $\bar{F}_l$ with the harmonic coefficients of $T(\Omega)$. In practice, we choose $T(\Omega)$ to be the gravity field due to instantaneous deglaciation if there had been no viscous relaxation. We refer to this choice of $T(\Omega)$ as $N^{\text{inst}}(\Omega)$, constructed by assuming isostatic equilibrium before deglaciation and uniform redistribution of the melted ice load over the ocean basins.

$\bar{F}_l$ describes the relation between the observed gravity field and N^{inst} . Because our initial distribution function, N^{inst} , is that corresponding to no relaxation, we should find that $\bar{F}_l < 1$ for all l . We also calculate the analogous quantity, $\bar{F}_l^M$, describing the relation between N^{inst} and predictions of the present-day glacial rebound gravity field from a suite of viscosity models (Fig. 2a). Comparison of data and models is done by using $\bar{F}_l$ because the models should be judged only after submitting them to the same filtering as the data.

To determine the model predictions of the present-day glacial rebound gravity field, we calculate the relaxation spectrum, τ_l^M , for each viscosity model assuming that the mantle can be described as a spherically symmetric newtonian fluid underlying a thin elastic lithosphere of thickness h (ref. 2). We consider three previously published viscosity models (LJN (ref. 20), HC (ref. 4) and FM (ref.

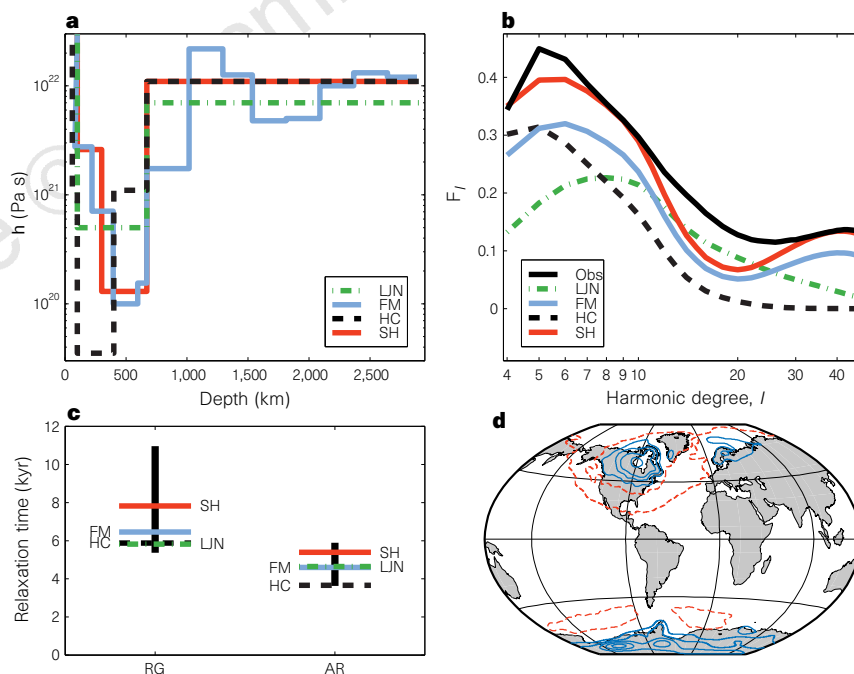


Figure 2 Viscosity model predictions versus observation. **a**, Viscosity, η , plotted against depth for models LJN, HC, FM and SH. All models use a 100-km-thick elastic surface layer, h , except model HC, which has $h = 60$ km. **b**, Observed and model predictions of $\bar{F}_l$. The estimates of ice thickness are taken from the ICE-3G deglaciation history²⁷. In estimating $\bar{F}_l$ we have made a simple correction to the observed gravity field, N^{obs} , for crustal thickness variations, ocean-continent differences and oceanic plate cooling²⁸. We also reference both N^{inst} and N^{obs} to the best-fitting ellipsoid. We make this choice because localization (using a window with characteristic length scale twice of that being considered) is not possible at $l = 2$. In particular, because N^{inst} is focused at the poles, it will have a significant correlation with the c_{20} term of N^{obs} , whether actually caused by glacial

rebound or by other processes, including convection or the delayed readjustment of the hydrostatic figure to changes in rotation rate¹⁰. Furthermore, because localization reduces spectral resolution, any difficulties at $l = 2$ can bleed into the neighbouring degrees. Therefore, we only consider here estimates of $\bar{F}_l$ for $l \geq 4$. **c**, Point relaxation times from RSL data at Richmond Gulf (RG) and Angerman River (AR). The vertical lines indicate the observed values with their errors²¹, and the horizontal lines indicate predictions from the indicated viscosity models. Model predictions are based on the best-fitting exponential to the elevations at AR and RG over the last 6 kyr. **d**, Predicted gravity field due to incomplete rebound with viscosity model SH; 6 mGal contour interval with solid and broken lines for contours ≤ 3 mGal and ≥ 3 mGal, respectively.

6)) as well as a new model (SH) that we developed specifically to fit the estimates of $\bar{F}_l$. The viscosity models are shown in Fig. 2a. Each τ_l^M spectrum is convolved with the deglaciation history to generate a predicted gravity model, N^M . The globally averaged local transfer function, $\bar{F}_l^M$, between N^{inst} and N^M can then be compared with $\bar{F}_l$ (Fig. 2b).

Model LNJ (ref. 20) is from a study of relative sea level (RSL) data in northwest Europe and uses no gravity data. This three-layer model considerably underestimates $\bar{F}_l$ at low l . The discrepancy indicates that either the longest wavelengths are poorly constrained by the RSL data or that the longest wavelengths of $\bar{F}_l$ are dominated by Hudson Bay and are not representative of the longest wavelengths in Fennoscandia.

Viscosity model HC has four viscous layers and is derived from a mantle flow calculation with a model of subducting lithosphere and lower-mantle seismic tomography as the buoyancy distribution⁴. Such flow models, used to infer viscosity distributions based on fits to the geoid, are sensitive only to relative viscosity variations and not to the absolute viscosity. To facilitate later comparisons we assign a lower-mantle viscosity to model HC of 1.1×10^{22} Pa s, which is taken directly from our new model SH. The salient feature of HC is the low (relative to the lower mantle) viscosities in the asthenosphere and transition zone. This viscosity variation was invoked to match the observed geoid anomalies over subduction zones⁴. This model consistently underestimates $\bar{F}_l$ and lacks the increase in $\bar{F}_l$ at high l .

Model FM (ref. 6) is from a joint inversion of RSL data at Angerman River (AR), Sweden, and Richmond Gulf (RG), Canada, together with a dynamic flow calculation done in a manner similar to that of HC, but by using a model of mantle heterogeneity inferred exclusively from seismic tomography. The dynamic flow component of the inversion attempts to fit only the $l = 2$ –8 components of the gravity field. This model has 13 viscosity layers. Of the models discussed so far, FM matches the shape of the observed $\bar{F}_l$ estimates the best, with maxima occurring at the same degrees. Despite the use of RSL data as calibration for the absolute viscosity, model FM also consistently underestimates $\bar{F}_l$. However, because model FM uses a global seismic tomography model of both the upper and lower mantle as input, it should be considered as a global model applying both to oceanic and continental regions.

Using forward modelling, we have found a three-layer viscosity model (SH) that adequately fits the $\bar{F}_l$ spectrum. The salient features of SH are the relatively high viscosity of the lower mantle (1.1×10^{22} Pa s) driven by the need to match the high values of $\bar{F}_l$ at low degrees, the low viscosity in the transition zone, and the 300 km thickness of the high-viscosity layer (2.6×10^{21} Pa s) in the upper mantle. A reduction in the thickness or magnitude of this upper-mantle high-viscosity layer degrades the fit to $\bar{F}_l$. The thick high-viscosity layer can also be found in model FM, and contrasts with models such as HC, which are based on flow models with the source of buoyancy in the upper mantle restricted to subduction zone environments. The logarithmic average of the upper-mantle viscosities in SH and FM is very near the upper-mantle value in model LNJ, which has a single layer for the entire sub-lithospheric upper mantle. All the viscosity models fail to match $\bar{F}_l$ near $l = 18$. Although the use of additional layers in SH could reduce this misfit, additional parameters are unwarranted given the simplicity of our modelling approach.

As an independent check on our viscosity model, we also consider the sea-level data used in constructing the LNJ and FM models. We use relaxation times for two sites chosen by Mitrovica²¹ and Forte and Mitrovica⁶ for AR and RG (Fig. 2c). These two sites were chosen because of their proximity to the centres of regions of deglaciation, thereby limiting their sensitivity to details in the model of deglaciation as well as limiting possible complications stemming from lateral variations in viscosity at the edges of the continents. The spatial relaxation times for AR, τ^{AR} , and RG, τ^{RG} , are ~ 4.7 and

7.6 kyr, respectively^{6,21}, although recent studies of other sea-level records in both of these areas indicate a spread of relaxation times²². These times are estimates at single points and represent the sum over all the wavelengths present at each location, thereby complicating any direct comparison with $\bar{F}_l$. We find that all the viscosity models fit within the range of acceptable values and that the differences between model prediction for τ^{AR} and τ^{RG} are entirely consistent with the differences in their estimates of $\bar{F}_l$. That model SH fits both the gravity and uplift data at Richmond Gulf but is near the upper bound of relaxation times for Angerman River might be indicative of variations in viscosity between the two sites.

When judging the success of a given viscosity model, the peak spatial gravity anomaly is a dangerous quantity to use because two different models can predict the same peak anomaly but with very different spectral contents (this is analogous to the non-uniqueness of the point RSL measurements). However, given a viscosity model, we can ask what portion of the total spatial gravity field is predicted to be due to incomplete rebound. Over Hudson Bay, model SH predicts a peak gravity anomaly of -24 mGal (Fig. 2d) and a peak geoid anomaly of -31 m. These values are just under half the total spatial gravity anomaly and over one-third of the total geoid anomaly in Hudson Bay. Over Fennoscandia, model SH predicts a peak gravity anomaly of about -10 mGal (Fig. 2d) and a peak geoid anomaly of about -6 m. Over Antarctica, model SH predicts a gravity anomaly of about -30 mGal and a geoid anomaly of about -26 m. These large anomalies probably reflect the relatively late deglaciation that is associated with the Antarctic ice sheet²³. The observed, predicted and residual gravity and geoid are shown in the Supplementary Information.

Viewed globally, the observed geoid and gravity fields have red spectra¹. This long-wavelength bias complicates any spatial analysis. Viewed locally, spectra vary greatly as a function of position—a reflection of the different geophysical processes that are active within a given region. To address this spatio-spectral complexity, we use a new localization method that renders more information than approaches that are either purely spatial or purely spectral. Because we are free to choose the extent of localization in either domain, our choice is non-unique. The spatio-spectral description presented here isolates the contribution to the gravity field from mantle convection, lithospheric-scale processes and incomplete glacial rebound. The rebound signature is greater than recent estimates and can be used to constrain models of mantle viscosity. We propose one such model that, although designed to explain the rebound portion of the gravity field, also matches observed RSL relaxation times. This model is characterized by a higher upper-mantle viscosity than viscosity models that explain the geoid highs associated with subducted slabs, consistent with a higher viscosity under shields than beneath oceans and continental margins. □

Received 27 January; accepted 23 September 1997.

1. Kaula, W. M. in *The Nature of the Solid Earth* (ed. Robertson, E. C.) 385–405 (McGraw-Hill, New York, 1972).
2. Walcott, R. I. Structure of the earth from glacio-isostatic rebound. *Annu. Rev. Earth. Planet. Sci.* **1**, 15–37 (1973).
3. Cathles, L. M. *The Viscosity of the Earth's Mantle* (Princeton Univ. Press, 1975).
4. Hager, B. H. & Clayton, R. W. in *Mantle Convection* (ed. Peltier, W. R.) 657–763 (Gordon and Breach, New York, 1989).
5. Peltier, W. R., Forte, A. M., Mitrovica, J. X. & Dziewonski, A. M. Earth's gravitational field: seismic tomography resolves the enigma of the Laurentian anomaly. *Geophys. Res. Lett.* **19**, 1555–1558 (1992).
6. Forte, A. M. & Mitrovica, J. X. New inferences of mantle viscosity from joint inversion of long-wavelength mantle convection and post-glacial rebound data. *Geophys. Res. Lett.* **23**, 1147–1150 (1996).
7. Pari, G. & Peltier, W. R. The free-air gravity constraint on subcontinental mantle dynamics. *J. Geophys. Res.* **101**, 28105–28132 (1996).
8. Hager, B. H. in *Glacial Isostasy, Sea-level and Mantle Rheology* (eds R. Sabadini, K. L. & Boschi, E.) 493–513 (NATO ASI Ser. C. vol. 334, Kluwer, Dordrecht, 1991).
9. Ricard, Y. & Vigny, C. Mantle dynamics with induced plate tectonics. *J. Geophys. Res.* **94**, 17543–17559 (1989).
10. O'Connell, R. J. Pleistocene glaciation and the viscosity of the mantle. *Geophys. J. R. Astron. Soc.* **23**, 299–327 (1971).
11. Mitrovica, J. X. & Peltier, W. R. Pleistocene deglaciation and the global gravity field. *J. Geophys. Res.* **94**, 13651–13671 (1989).
12. Simons, M. thesis, MIT, Cambridge, Massachusetts (1996).

13. Simons, M., Solomon, S. C. & Hager, B. H. Localization of gravity and topography: constraints on the tectonics and mantle dynamics of Venus. *Geophys. J. Int.* **131**, 24–44 (1997).
14. Hager, B. H. Subducted slabs and the geoid: constraints on mantle rheology and flow. *J. Geophys. Res.* **89**, 6003–6015 (1984).
15. Masters, G., Johnson, S., Laske, G. & Bolton, H. A shear-velocity model of the mantle. *Phil. Trans. R. Soc. Lond. A* **354**, 1385–1411 (1996).
16. Van Bemmelen, R. W. & Berlage, H. P. Versuch einer mathematischen behandlung geotektonischer bewegung unter besonderer berucksichtigung der undationstheorie. *Beitr. Geophys.* **43**, 19–55 (1935).
17. Haskell, N. A. The viscosity of the asthenosphere. *Am. J. Sci.* **33**, 22–28 (1937).
18. Forte, A. M. & Peltier, W. R. Viscous flow models of global geophysical observables. 1. Forward problems. *J. Geophys. Res.* **96**, 20131–20159 (1991).
19. Forte, A. M., Woodward, R. L. & Dziewonski, A. M. Joint inversions of seismic and geodynamic data for models of three-dimensional mantle heterogeneity. *J. Geophys. Res.* **99**, 21857–21877 (1994).
20. Lambeck, K., Johnston, P. & Nakada, M. Glacial rebound and sea-level change in northwestern Europe. *Geophys. J. Int.* **103**, 451–468 (1990).
21. Mitrovica, J. X. Haskell [1935] revisited. *J. Geophys. Res.* **101**, 555–569 (1996).
22. Mitrovica, J. X. & Forte, A. M. Radial profile of mantle viscosity—results from the joint inversion of convection and postglacial rebound observables. *J. Geophys. Res.* **102**, 2751–2769 (1997).
23. Nakada, M. & Lambeck, K. Glacial rebound and relative sea-level variations: a new appraisal. *Geophys. J. R. Astron. Soc.* **90**, 171–224 (1987).
24. Nerem, R. S. *et al.* Gravity model development for TOPEX/Poseidon: joint gravity models 1 and 2. *J. Geophys. Res.* **99**, 24421–24447 (1994).
25. Nerem, R. S., Jekeli, C. & Kaula, W. M. Gravity field determination and characteristics: retrospective and prospective. *J. Geophys. Res.* **100**, 15053–15074 (1995).
26. Wessel, P. & Smith, W. H. F. New version of the Generic Mapping Tools released. *Eos* **76**, 329 (1991).
27. Tushingham, A. M. & Peltier, W. R. Ice-3g: a new global model of late Pleistocene deglaciation based upon geophysical predictions of post-glacial relative sea-level change. *J. Geophys. Res.* **96**, 4497–4523 (1991).
28. Hager, B. H. Global isostatic geoid anomalies for plate and boundary layer models of the lithosphere. *Earth Planet. Sci. Lett.* **63**, 97–109 (1983).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements. We thank M. Fang, M. Gurnis and S. Zhong for constructive discussions, as well as J. X. Mitrovica for a thorough review. This work was supported by NASA.

Correspondence should be addressed to M.S. (e-mail: simons@gps.caltech.edu).

Cosmopolitanism among Gondwanan Late Cretaceous mammals

David W. Krause*, G. V. R. Prasad†, Wighart von Koenigswald‡, Ashok Sahni§ & Frederick E. Grinell

* Department of Anatomical Sciences, and † Departments of Anthropology and Anatomical Sciences, State University of New York at Stony Brook, Stony Brook, New York 11794, USA

† Department of Geology, University of Jammu, Jammu 180 004, India

‡ Institute of Palaeontology, University of Bonn, Nussallee 8, D-53115, Bonn, Germany

§ Centre of Advanced Study in Geology, Panjab University, Chandigarh 160 014, India

Consistent with geophysical evidence for the breaking up of Pangaea, it has been hypothesized that Cretaceous vertebrates on progressively isolated landmasses exhibit generally increasing levels of provincialism^{1–3}, with distinctly heightened endemism occurring at the beginning of the Late Cretaceous⁴. The Cretaceous fossil record from the southern supercontinent of Gondwana has been much too poor to test this hypothesis with regards to mammals (Fig. 1). Early Cretaceous mammals are known only from isolated sites in Argentina⁵, Australia^{6,7}, Cameroon^{8,9} and Morocco¹⁰. Apart from several occurrences in South America¹¹, knowledge of Late Cretaceous Gondwanan mammals is limited to a single site in India that previously yielded a few specimens of placental mammals^{12,13}, and a site in Madagascar that previously yielded only one indeterminate tooth fragment¹⁴. Here we report the occurrence of a highly specialized and distinctive group of extinct mammals, the Sudamericiidae (Gondwanatheria), in the Late Cretaceous of Madagascar and India. These new records

comprise the first evidence of gondwanatheres outside South America and the first indication of cosmopolitanism among Late Cretaceous Gondwanan mammals. Antarctica may have served as an important Cretaceous biogeographic link between South America and Indo-Madagascar.

The Gondwanatheria is a group of multituberculate^{15,16} or multituberculate-like¹⁷ mammals previously known only from the Late Cretaceous and Palaeocene of Argentina. Apart from two tentatively referred, fragmentary dentaries (one edentulous and the other bearing a single tooth), Gondwanatheria is based on isolated teeth. The gondwanathere teeth described here are the first identifiable, pre-Late Pleistocene specimens of non-marine mammals known from Madagascar and the first remains of non-placental mammals from the Cretaceous of the Indian subcontinent. The specimens from Madagascar were recovered during joint expeditions by the State University of New York, Stony Brook, and the Université d'Antananarivo in 1995 and 1996 to the continental Upper Cretaceous Maevarano Formation, which, in addition to mammals, has recently yielded a diversity of vertebrate taxa¹⁸. The gondwanathere specimen from India was discovered in 1989 in sediments of the Upper Cretaceous Deccan Intertrappean sequence^{12,13}, but was not identified as that of a gondwanathere until the specimens from Madagascar were examined. None of the lower taxa of mammals from the Late Cretaceous were previously known to be spread across South America, India and Madagascar.

Mammalia Linnaeus 1758

?Allotheria Marsh 1880

Gondwanatheria Mones 1987

Sudamericiidae Scillato-Yané & Pascual 1984

Lavanify miolaka gen. et sp. nov.

Etymology. *Lavanify* (la-va-NEE-fee; Malagasy), long tooth; *miolaka* (MYOU-la-ka; Malagasy), curved; in reference to the shape of the cheek-teeth.

Holotype. Université d'Antananarivo (UA) 8653, well-preserved cheek-tooth (Fig. 2a, d).

Referred specimen. Field Museum of Natural History (FMNH) PM 59520, fragmentary cheek-tooth (Fig. 2b).

Localities and horizon. Holotype from locality MAD96-01 and referred specimen from locality MAD93-35, uppermost white sandstone unit of Upper Cretaceous (?Campanian) Maevarano Formation, Mahajanga Basin, near village of Berivotra, northwestern Madagascar.

Diagnosis. The teeth of *Lavanify* differ from those of the only previously known sudamericiid genera *Gondwanatherium* and *Sudamerica* in possessing prominent and continuous inter-row sheets of interprismatic matrix in dental enamel and at least one cheek-tooth position that has a single, V-shaped dentine island and lacks enamel on one side of the crown. *Lavanify* further differs from *Gondwanatherium* in having cheek-teeth with vertical furrows that extend to the base of the crown and onto the root.

Description. UA 8653, the holotype specimen, is a molariform, hypsodont cheek-tooth (Fig. 2a). Its preserved height is 11.2 mm, and what are interpreted to be its length and width are 3.4 and 3.2 mm, respectively. The crown, as determined from the distribution of enamel, comprises about 85% of the tooth's height. UA 8653 is strongly curved along its height and worn flat on its occlusal surface. In occlusal view, the worn surface consists of a broad, V-shaped dentine island surrounded by enamel (except along one edge, where breakage has occurred). The indentation of the V is formed by a vertical furrow that extends through the entire height of the tooth and is filled with cementum. Enamel is clearly absent from one side of the crown, where two distinct, vertical enamel–dentine edges are evident. UA 8653 possesses small, circular enamel prisms aligned in rows, which are separated by prominent and continuous bands of interprismatic material (Fig. 2d).

FMNH PM 59520 (Fig. 2b) is tentatively referred to *L. miolaka*

because, like UA 8653, it is large (preserved height, 9.8 mm; length and width cannot be determined), molariform, hypsodont, curved along its height (although slightly less so than UA 8653), worn flat on its occlusal surface, and possesses rows of small, circular enamel prisms separated by prominent inter-row sheets. Unlike UA 8653, FMNH PM 59520 possesses an infundibulum that invaginates deeply from the occlusal surface and contains cementum. A cementum-filled furrow or a second, incompletely preserved infundibulum is present as well, but breakage precludes a determination of which of the two possible conditions exist. The presence of an infundibulum and the lesser degree of curvature suggest that FMNH PM 59520 represents a different tooth position from the holotype. In the absence of conclusive evidence to the contrary, and in the light of the considerable range in morphological variation in molariform tooth positions among other gondwanatheres species^{1,15,16}, UA 8653 and FMNH PM 59520 are provisionally considered to represent the same species.

The isolated nature of the two specimens of *Lavanify miolaka*, as well as poor knowledge of all other known gondwanatheres, precludes a confident determination of whether the specimens represent upper or lower teeth, or whether they are molars or molariform premolars.

Sudamericidae incertae sedis
Genus and species indeterminate

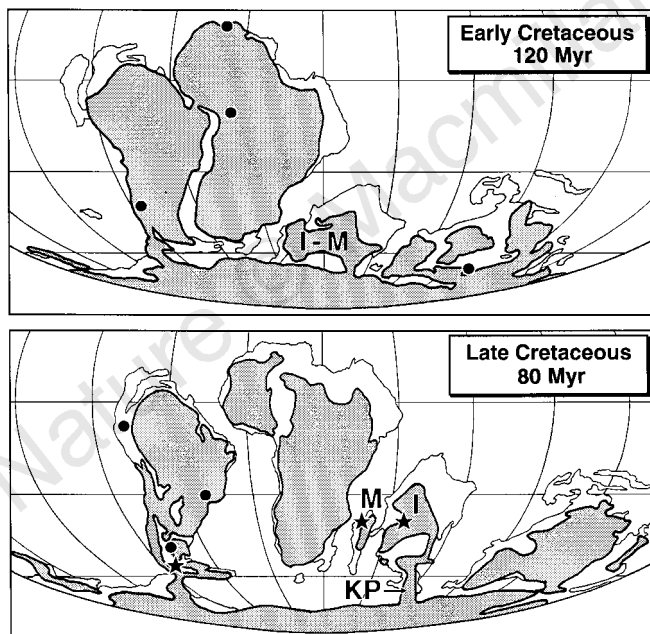


Figure 1 Maps showing distribution of Gondwanan landmasses and known mammal localities in the Early (top) and Late Cretaceous (bottom). Shaded areas indicate the distribution of subaerially exposed land (adapted from ref. 26); dots indicate mammal localities; stars indicate mammal localities that have yielded gondwanatheres. Madagascar (M), with the Indian subcontinent (I) attached to its current eastern margin, separated from Africa ~165 Myr ago and attained its current position relative to the mainland ~124 Myr (refs 23, 25, 29, 30). Strike-slip motion between the Indian subcontinent and Madagascar began ~135 Myr (ref. 25), as India rifted from Antarctica, but they remained in proximity until ~88 Myr (ref. 22), several million years before deposition of the Maevarano Formation¹⁸. Indo-Madagascar (I-M) and eastern Antarctica were connected until at least ~120 Myr (ref. 25), and possibly as late as 80 Myr (ref. 26), across the Kerguelen Plateau (KP). The Antarctic Peninsula remained very close to, or maintained contact with, South America in the Late Cretaceous and into the early Tertiary^{23,24,26}.

Referred specimen. Vertebrate Palaeontology Laboratory, Jammu University, Naskal Intertrappean Mammal (VPL/JU/NKIM) 25, fragmentary cheek-tooth (Fig. 2c, e).

Locality and horizon. Naskal Locality, Andhra Pradesh, central India, in Upper Cretaceous (late Maastrichtian) Deccan Intertrappean sequence.

Description. VPL/JU/NKIM/25 is too incomplete to assign to a particular genus and species, but is complete enough to demonstrate clearly that it possesses a range of characteristics unique to the Sudamericidae among Cretaceous mammals. Although smaller (about 6 mm in height), it resembles the cheek-tooth specimens of known sudamericids, including those of *Lavanify miolaka*, in being molariform, hypsodont, and curved along its height, as well as in possessing a flat occlusal surface and an infundibulum that invaginates far into the crown (Fig. 2c). It also exhibits small, circular prisms aligned in rows that are separated by well-developed inter-row sheets of interprismatic matrix (Fig. 2e). Like *Lavanify*

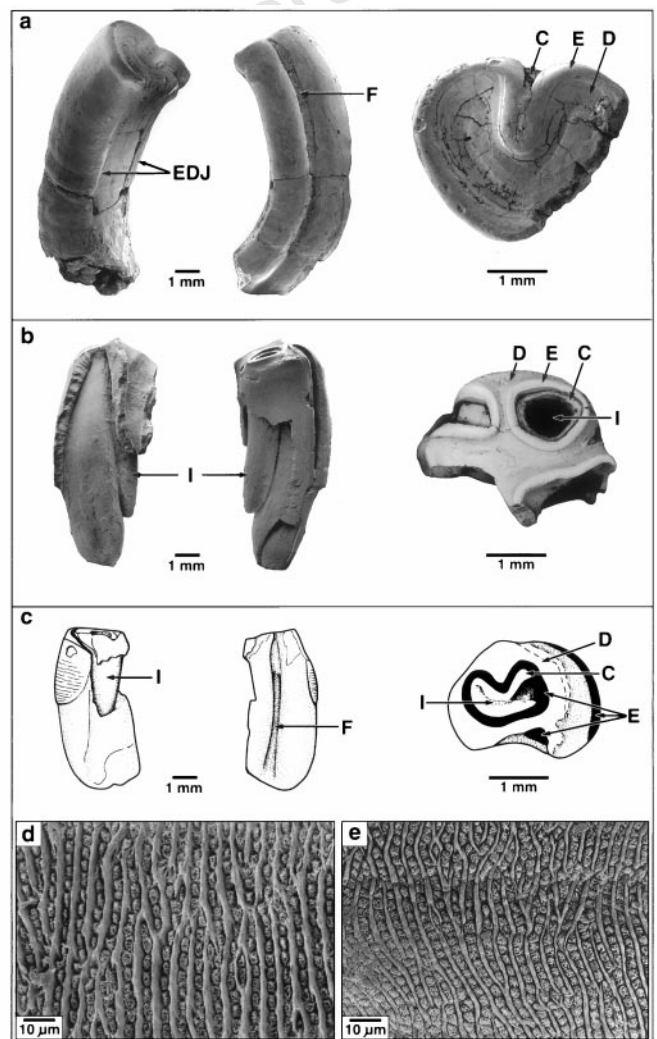


Figure 2 Sudamericids from Madagascar and India. **a, b**, Scanning electron micrographs of cheek-teeth of *Lavanify miolaka* from the Upper Cretaceous Maevarano Formation, Mahajanga Basin, northwestern Madagascar. **a**, UA 8653, holotype, in side (left, middle) and occlusal (right) views. **b**, FMNH PM 59520 in side (left, middle) and occlusal (right) views. **c**, Drawings of cheek-tooth (VPL/JU/NKIM/25) of unnamed Indian sudamerid from the Upper Cretaceous Deccan Intertrappean sequence, India, in side (left, middle) and occlusal (right) views. **d, e**, Scanning electron micrographs of tangential sections of enamel microstructure in *L. miolaka* (UA 8653) and unnamed Indian sudamerid (VPL/JU/NKIM/25), respectively. C, cementum; D, dentine; E, enamel; EDJ, enamel-dentine junction; F, furrow; I, infundibulum.

and *Sudamerica*, the Indian form possesses a prominent, cementum-filled vertical furrow that extends throughout the entire height on the side of the crown. It cannot be demonstrated that enamel is absent on one side of the crown, as in UA 8653, the holotype of *L. miolaka*.

The specimens of *Lavanify* from Madagascar and the unnamed Indian form exhibit several features diagnostic of Gondwanatheria, which were previously known only from Argentina and composed of the ferugliotheriid *Ferugliotherium* (Late Cretaceous, Campanian) and the sudamericids *Gondwanatherium* (Late Cretaceous, Campanian) and *Sudamerica* (early Palaeocene)^{1,15–17}. *Lavanify* and the Indian form are more derived than other gondwanatheres and are tentatively linked with each other on the basis of the shared possession of prominent and continuous inter-row sheets of interprismatic matrix in the enamel (Fig. 3). Although the interprismatic matrix is thick in *Gondwanatherium* and *Sudamerica*^{15,19}, and is, in places, organized into indistinct inter-row sheets¹⁹, it anastomoses around prisms and the sheets are not nearly as well-organized into continuous separate bands as they are in *Lavanify* and the Indian form. However, because this level of differentiation has been recorded at varying depths (albeit of much thicker enamel) in single molars of extant placental mammal species (caprine bovids)²⁰, we caution that additional samples must be discovered and analysed to assess the range of variation in the development of inter-row sheets within gondwanathere species and, thereby, to validate this synapomorphy. *Lavanify* is considered autapomorphic in possessing at least one cheek-tooth position with a single, V-shaped dentine island and the absence of enamel on one side of the crown.

Within the context of current geophysical models for the sequence of fragmentation of Gondwana (Fig. 1), the discovery of sudamericids in the Late Cretaceous of Madagascar and India, in addition to their occurrence in Argentina, has several profound biogeographic implications. First, the Sudamericiidae were distributed much more broadly among Gondwanan landmasses during the Late Cretaceous than previously known. Until now it was not known whether the most well-sampled mammalian fauna from the

Late Cretaceous of Gondwana, from the Los Alamos Formation in Patagonia and which consists of non-tribosphenic forms only, was wholly endemic to Patagonia, South America or Gondwana^{1,15–17}. The presence of sudamericids in both Madagascar and India, in addition to their occurrence in South America, demonstrates clearly that at least some Late Cretaceous mammalian taxa were broadly distributed on Gondwana.

Second, the mammalian fauna from the Late Cretaceous of Madagascar was not highly endemic. The modern mammalian fauna of Madagascar is wholly endemic at the species level and includes several suprageneric taxa that are unique to the island (Lemuroidea, Tenrecinae, Nesomyinae, Galidiinae, Euplerinae and Cryptoproctinae)¹⁸. Because *Lavanify* does not bear a close phylogenetic relationship to any of the modern and recently extinct orders of mammals from Madagascar, it contributes no positive evidence to help solve the mystery of their biogeographic origins.

Third, the mammalian fauna from the Late Cretaceous of the Indian subcontinent, like other elements of the biota²¹, has components that are shared with other Gondwanan landmasses. The only other mammalian taxa described from the Indian Late Cretaceous, the palaeoryctid eutherians *Deccanolestes hislopi* and *D. robustus*, have been suggested to have affinities with taxa of Laurasian origin^{12,13}. The tentative phylogenetic placement of the Indian gondwanathere and *Lavanify* as sister taxa is consistent with geophysical evidence indicating that the Indian subcontinent separated from Madagascar approximately 88 million years ago (Myr)²².

Finally, as well as occurring in South America, Madagascar and the Indian subcontinent, sudamericids must have been present on some intervening landmass(es) in the Late Cretaceous or earlier. The three most likely hypotheses to explain the disjunct distribution of sudamericids in Argentina, Madagascar and India are that sudamericids were: found across the whole of Gondwana; also present in Africa alone; or also present in Antarctica alone. Within known geophysical constraints, and acknowledging that Late Cretaceous vertebrate faunas of Gondwana are generally poorly known, the hypothesis that sudamericids were also present in Antarctica seems the most likely. South America was close to or contiguous with the Antarctic Peninsula in the Late Cretaceous and Indo-Madagascar was close to or contiguous with Antarctica until the late Early Cretaceous^{23–25}. A revised palaeogeographic reconstruction of this region²⁶ suggests a connection between Antarctica and Indo-Madagascar across the Kerguelen Plateau that lasted well into the Late Cretaceous (possibly as late as 80 Myr), as shown in Fig. 1. In contrast, a narrow sinuous seaway surrounding southern Africa developed in the late Early Cretaceous and progressively widened in the Late Cretaceous, isolating Africa from South America and Antarctica^{23,26}. Furthermore, the hypotheses that sudamericids were pan-Gondwanan or were also present in Africa alone predict that either they would have evolved by the Late Jurassic, before the breaking-up of Gondwana in general and before Madagascar rifted from Africa in particular, or that they crossed the Mozambique Channel, which is over 400 km wide at its narrowest point, some time later. Basal gondwanatheres may have originated on, and been restricted to, Gondwana by the Late Jurassic. Sudamericids, however, are highly derived, and it therefore seems unlikely that they existed in the Late Jurassic; whether or not they crossed the Mozambique Channel at a later time is currently untestable. Finally, support for Late Cretaceous biotic connections among South America, Antarctica, Madagascar and the Indian subcontinent can also be derived from other terrestrial vertebrate taxa, particularly dinosaurs. For instance, the derived abelisaurid ceratosaurian theropods *Magungasaurus* and *Indosuchus*, from the Late Cretaceous of Madagascar and India, respectively, are sister taxa and are closely related to taxa from the Cretaceous of Argentina²⁷. In contrast, although fragmentary remains of purported abelisaurids have been recently reported from the Early Cretaceous of Africa²⁸ (as well as from the Late Cretaceous of Europe), the Late Cretaceous theropod fauna of

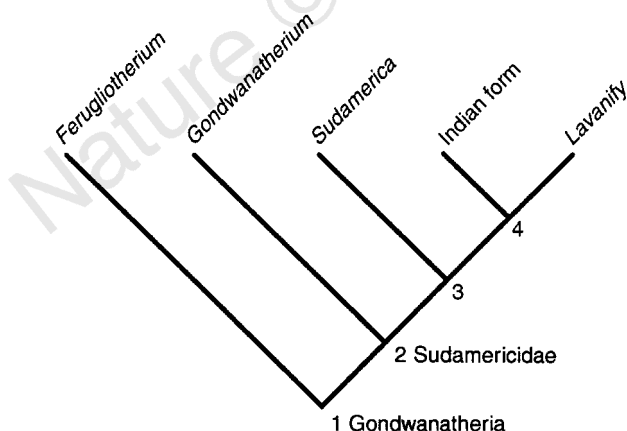


Figure 3 Cladogram showing placement of *Lavanify* and the unnamed Indian sudamericid within Gondwanatheria. The unambiguous synapomorphies diagnosing the nodes are: 1 (Gondwanatheria), prominent transverse ridges and furrows on molariform cheek-tooth crowns, incisors enlarged and procumbent with restricted band of enamel, small and circular enamel prisms; 2 (Sudamericiidae), large size, cheek-teeth hypsodont, cheek-teeth curved through height, prominent furrows present on sides of molariform cheek-teeth, infundibula that extend far cervically into molariform cheek-tooth crowns present, cementum present on sides of cheek-tooth crown and inside infundibula, at least one cheek-tooth position with obliquely oriented occlusal plane; 3, presence of furrows on sides of molariform cheek-tooth extending far towards midline of tooth and cervically onto root; and 4, enamel with continuous and prominent inter-row sheets of interprismatic matrix.

Africa appears to have been dominated by tetanurans, including carcharodontosaurids, spinosaurids and coelurosaur⁴.

As the Cretaceous fossil record of vertebrates on southern landmasses improves, so too will the ability to test the biogeographic conclusions outlined above. Crucial to these tests will be the recovery of mammalian and other vertebrate fossils from the Late Cretaceous of northern South America, mainland Africa, and Australia, to determine whether or not gondwanatheres and other vertebrate taxa were restricted to a high-latitude route involving only Argentina, Antarctica, Madagascar and the Indian subcontinent, or whether they were distributed even more broadly on Gondwana. At the very least, however, our evidence demonstrates that at least some elements of Late Cretaceous vertebrate faunas were broadly shared among Gondwanan landmasses. □

Received 21 July; accepted 19 September 1997.

1. Bonaparte, J. New Late Cretaceous mammals from the Los Alamos Formation, northern Patagonia. *Natl Geogr. Res.* **6**, 63–93 (1990).
2. Le Loeuff, J. The Campano-Maastrichtian vertebrate faunas from southern Europe and their relationships with other faunas in the world; palaeobiogeographical implications. *Cretaceous Res.* **12**, 93–114 (1991).
3. Russell, D. A. China and the lost worlds of the dinosaurian era. *Hist. Biol.* **10**, 3–12 (1995).
4. Sereno, P. C. *et al.* Predatory dinosaurs from the Sahara and Late Cretaceous faunal differentiation. *Science* **272**, 986–991 (1996).
5. Bonaparte, J. & Rougier, G. Mamíferos del Cretácico Inferior de Patagonia. *Mem. IV Congr. Latinoamericano de Paleontol., Bolivia* **1**, 343–359 (1987).
6. Archer, M., Flannery, T. F., Ritchie, A. & Molnar, R. E. First Mesozoic mammal from Australia—an early Cretaceous monotreme. *Nature* **318**, 363–366 (1985).
7. Flannery, T. F., Archer, M., Rich, T. H. & Jones, R. A new family of monotremes from the Cretaceous of Australia. *Nature* **377**, 418–420 (1995).
8. Jacobs, L. L. *et al.* Mammal teeth from the Cretaceous of Africa. *Nature* **336**, 158–160 (1988).
9. Brunet, M. *et al.* Nouveaux mammifères du Crétacé inférieur du Cameroun, Afrique de l'Ouest. *C. R. Acad. Sci.* **310**, 1139–1146 (1990).
10. Sigogneau-Russell, D., Monbaron, M. & Russell, D. E. Découverte de Mammifères dans le Mésozoïque moyen d'Afrique. *C. R. Acad. Sci.* **307**, 1045–1050 (1988).
11. Marshall, L. G. & Sempere, T. in *Biological Relationships Between Africa and America* (ed. Goldblatt, P.) 329–392 (Yale Univ. Press, New Haven, CT, 1993).
12. Prasad, G. V. R. & Sahni, A. First Cretaceous mammal from India. *Nature* **332**, 638–640 (1988).
13. Prasad, G. V. R., Jaeger, J. J., Sahni, A., Gheerbrant, E. & Khajuria, K. K. Eutherian mammals from the Upper Cretaceous (Maastrichtian) Intertrapean beds of Naskal, Andhra Pradesh, India. *J. Vert. Paleontol.* **14**, 260–277 (1994).
14. Krause, D. W. *et al.* Late Cretaceous mammals. *Nature* **368**, 298 (1994).
15. Krause, D. W., Kielan-Jaworowska, Z. & Bonaparte, J. F. *Ferugliotherium* Bonaparte, the first known multituberculata from South America. *J. Vert. Paleontol.* **12**, 351–376 (1992).
16. Krause, D. W. & Bonaparte, J. F. Superfamily Gondwanatherioidea: A previously unrecognized radiation of multituberculata mammals in South America. *Proc. Natl Acad. Sci. USA* **90**, 9379–9383 (1993).
17. Pascual, R., Goin, F. J., Ortiz-Jaureguizar, E., Carlini, A. & Prado, J. L. *Ferugliotherium* and *Sudamerica*, Multituberculata and Gondwanatheria: One more evolutionary process occurred in isolation. *Ameghiniana* **30**, 334–335 (1993).
18. Krause, D. W., Hartman, J. H. & Wells, N. A. in *Natural Change and Human Impact in Madagascar* (eds Goodman, S. M. & Patterson, B. D.) 3–43 (Smithsonian Inst. Press, Washington DC, 1997).
19. Sigogneau-Russell, D., Bonaparte, J. F., Frank, R. M. & Escibano, V. Ultrastructure of dental hard tissues of *Gondwanatherium* and *Sudamerica* (Mammalia, Gondwanatheria). *Lethaia* **24**, 27–38 (1991).
20. Grine, F. E., Fosse, G., Krause, D. W. & Jungers, W. L. Analysis of enamel ultrastructure in archaeology: the identification of *Ovis aries* and *Capra hircus* dental remains. *J. Archaeol. Sci.* **13**, 579–595 (1986).
21. Sahni, A. Cretaceous-Paleocene terrestrial faunas of India: Lack of endemism during drifting of the Indian plate. *Science* **226**, 441–443 (1984).
22. Storey, M. *et al.* Timing of hot spot-related volcanism and the breakup of Madagascar and India. *Science* **267**, 852–855 (1995).
23. Lawver, L. A., Gahagan, L. M. & Coffin, M. F. The development of paleoseaways around Antarctica. *Antarctic Res. Ser.* **56**, 7–30 (1992).
24. Shen, Y. A paleoisthmus between southern South America and Antarctic Peninsula during Late Cretaceous and Early Tertiary. *VII Int. Symp. Antarctic Earth Sciences Abstr.* **345**, 345 (1995).
25. Roeser, H. A., Fritsch, J. & Hinz, K. The development of the crust off Dronning Maud Land, East Antarctica. *Geol. Soc. Spec. Publ.* **108**, 243–264 (1996).
26. Hay, W. W. *et al.* in *The Evolution of Cretaceous Ocean/Climate Systems* (eds Barrera, E. & Johnson, C.) (Geological Society of America, special publication, in the press).
27. Sampson, S. D., Krause, D. W., Dodson, P. & Forster, C. A. The premaxilla of *Magungasaurus* (Dinosauria: Theropoda) with implications for Gondwanan paleobiogeography. *J. Vert. Paleontol.* **16**, 601–605 (1996).
28. Russell, D. A. Isolated dinosaur bones from the Middle Cretaceous of the Tafilalet, Morocco. *Bull. Mus. Natl Hist. Nat. Paris* **18**, 349–402 (1996).
29. Smith, A. G., Smith, D. G. & Funnell, B. M. *Atlas of Mesozoic and Cenozoic Coastlines* (Cambridge Univ. Press, 1994).
30. Coffin, M. F. & Rabinowitz, P. D. The Mesozoic East African and Madagascan conjugate continental margins. *Am. Assoc. Petrol. Geol. Mem.* **53**, 207–246 (1992).

Acknowledgements. We thank B. Rakotosamimanana, P. Wright, B. Andriamihaja and the Institute for the Conservation of Tropical Environments staff for help with fieldwork in Madagascar; members of the 1995 and 1996 field crews for their efforts; J. Ratsimbazafy for Malagasy translations; C. Forster, R. Fox, Z. Kielan-Jaworowska, R. Presley, T. Rich, G. Rougier and S. Sampson for discussion and/or review of earlier drafts of the manuscript; W. Hay for permission to cite unpublished work; and L. Betti-Nash and M. Stewart for assistance with the figures. This work was funded by the National Science Foundation, the National Geographic Society, and the Deutsche Forschungsgemeinschaft.

Correspondence and requests for materials should be addressed to D.W.K. (e-mail: dkrause@mail.som.sunysb.edu).

Biodiversity enhances ecosystem reliability

Shahid Naeem & Shubin Li

Department of Ecology, Evolution and Behavior and Center for Community Genetics, University of Minnesota, 100 Ecology Building, 1987 Upper Buford Circle, St Paul, Minnesota 55108, USA

Biodiversity may represent a form of biological insurance against the loss or poor performance of selected species¹. If this is the case, then communities with larger numbers of species should be more predictable with respect to properties such as local biomass². That is, larger numbers of species should enhance ecosystem reliability, where reliability refers to the probability that a system will provide a consistent level of performance over a given unit of time³. The validity of this hypothesis has important ecological, management and economic implications given the large-scale substitution of diverse natural ecosystems with less diverse managed systems⁴. No experimental evidence, however, has supported this hypothesis⁵. To test this hypothesis we established replicate microbial microcosms with varying numbers of species per functional group. We found that as the number of species per functional group increased, replicate communities were more consistent in biomass and density measures. These results suggest that redundancy (in the sense of having multiple species per functional group^{6–8}) is a valuable commodity, and that the provision of adequate redundancy may be one reason for preserving biodiversity.

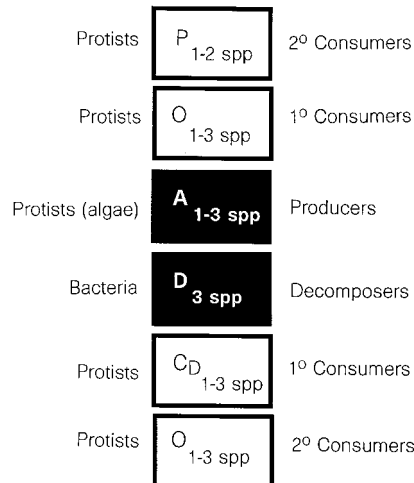


Figure 1 Experimental design. Each microcosm contained two core trophic (functional) groups: autotrophs (A) and decomposers (D). Additional functional groups consisted of species classified as either primary (1°) or secondary (2°) consumers of either A or D. Consumers were all protists species with 1–3 species per functional group, with the exception of top predators (P). Species were chosen at random from the pools of available species (listed below). Because several consumer species were omnivorous (capable of feeding on more than one trophic level) they were classified as omnivores (O). Species in A were *Chlamydomonas reinhardtii*, *Trachelomonas* sp. and *Euglena gracilis*. Species in D were *Serratia marcescens*, *Bacillus subtilis* and *Bacillus cereus*. Consumers of decomposers, C_D, were the ciliate species *Colpidium striatum*, *Spirostomum ambiguum*, *Vorticella* sp., *Stentor* sp. and the flagellate, *Chilomonas paramecium*. Species in O were the ciliates *Euplotes eurystomus*, *Paramecium aurelia*, *Pa. caudatum*, *Pa. multimicronucleatum*, *Blepharisma* sp. and the sarcodinid, *Amoeba proteus*. Species in P consisted of the sarcodinid *Pelomyxa carolinensis* and the ciliate *Didinium nasutum*.

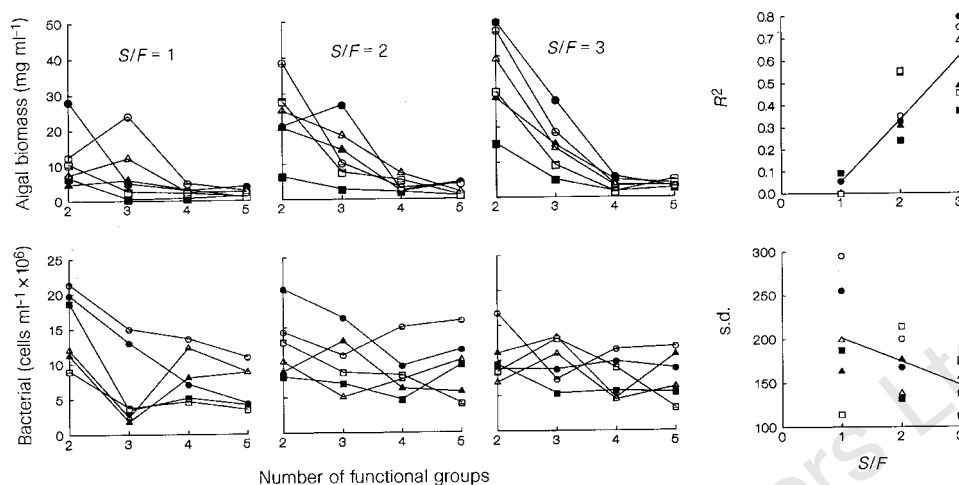


Figure 2 Ecosystem reliability and S/F . Top, algal biomass is plotted against the number of functional (trophic) groups (see Fig. 1). Each plot shows results for microcosms representing different S/F values established under three nutrient and two light conditions (see below for explanation of symbols). Each point represents the mean of four replicates. The right-most plot shows the R^2 of a linear regression fitted to the natural log of the autotrophic biomass as a function of number of functional groups in a microcosm in relation to S/F irrespectively of light and nutrient conditions. The line represents a fitted least-squares model illustrating that autotrophic biomass becomes increasingly predictable as a

function of F as S/F increases. Bottom, bacterial density plotted against the number of functional (trophic) groups as in the top figure. The right-most plot shows the standard deviation (s.d.) of bacterial densities among replicate microcosms as a function of S/F , showing that variation in bacterial densities declined with increasing S/F regardless of light and nutrient conditions. The line represents a fitted least-squares model. Open symbols, high light; filled symbols, low light; circles, high nutrients (full concentration); triangles, intermediate nutrients (1/2 concentration); squares, low nutrients (1/4 concentration).

The ecological basis for biological insurance is compensatory growth², a widely observed process in which one species within a functional group increases in response to the reduction or loss of another in the same functional group^{9–12}. Functional groups, or groups of ecologically equivalent species, are operationally defined by the ecosystem properties being measured¹³. For multitrophic systems (that is, containing autotrophs such as plants or phytoplankton, decomposers and consumers such as herbivores and predators), functional groups may be operationally defined as trophic groups^{14,15}. Thus, ecosystem reliability should increase when the number of species (S) per functional group (F), or S/F , increases because larger numbers of species per functional group increases the probability that compensatory growth will occur³.

We tested this hypothesis by directly manipulating diversity in model ecosystems. We constructed 318 microbial microcosms, each of which contained inorganic and organic nutrients, photoautotrophs and decomposers; the critical components of terrestrial and aquatic ecosystems¹⁶. All environmental conditions were held constant, but we varied nutrients (3 levels), light (2 levels), S/F (three levels) and F (4 levels) (Fig. 1). We measured biomass and density of photoautotrophs and decomposers (bacteria) as response variables at the end of the experiment (57 days). These response variables served as complex indices of the end point of assimilation, immobilization, photosynthesis, respiration and decomposition, the principal processes of natural ecosystems¹⁶, over the duration of the experiment. Species per functional group were chosen at random from species pools. Local extinction occurred, as is common for microcosms in the absence of colonization¹⁷, and this allowed for compensatory growth. If S/F increases ecosystem reliability, then the response variables should show greater predictability or lower variance irrespective of light or nutrient levels.

These model ecosystems proved to be surprisingly more predictable as S/F increased. In general, algal biomass declined with increasing F for all levels of S/F . The relationship between algal biomass and F was weak, however, when $S/F = 1$, but increased significantly as S/F increased (Fig. 2). Bacterial densities showed no response to variation in F or S/F (Fig. 2). The standard deviation

(s.d.) of decomposer density among microcosms, however, showed a strong decline as S/F increased (Fig. 2).

The increased reliability generated by higher levels of S/F was independent of light and nutrient levels. A full factorial three-way analysis of variance indicated that nutrients did have significant effects on mean algal biomass and mean bacterial densities ($P < 0.05$), whereas light levels had no significant effects on either response variable. These analyses also indicated that S/F had significant effects on mean algal biomass and extinction rates ($P < 0.01$). Autotrophic biomass was lower (Fig. 2) and extinction higher in microcosms with higher F . Thus, S/F affected both the mean and variance of ecosystem biomass. Consequently, after local extinction, compensatory growth by remaining species within functional groups provided more predictable levels of biomass and lower variance in bacterial densities in higher-diversity microcosms regardless of nutrient and light conditions.

The results provide support for the biological insurance hypothesis for biodiversity. The design of these microcosms focused on average or global ecosystem response to variation in S/F . Individually, or on a local rather than global scale, species-specific interaction strengths are likely to provide better predictions for ecosystem response to variation in biodiversity¹⁸. On average, however, for a collection of increasingly depauperate ecosystems in which S/F declines, ecosystems may become increasingly less reliable in terms of the goods, services¹⁹ and economic values²⁰ they provide. □

Methods

Microcosms. Microcosms consisted of disposable, sterile 100-mm-diameter $\times$ 25-mm-deep polystyrene Petri dishes each filled with 50 ml of growth medium.

Treatments. Growth medium was prepared as in ref. 21. 'Full' nutrient treatment levels represented this growth medium in an undiluted state whereas '1/2' and '1/4' nutrients represented dilutions by 1/2 and 1/4 using sterile water, respectively. Light was set at two levels by manipulating the number of lamps in chambers; 'low' treatment levels equalled a mean of 31.09 (s.d. = 12.58) $\mu\text{mol m}^{-2} \text{s}^{-1}$ and 'high' equalled a mean of 52.79

(s.d. = 11.06) $\mu\text{mol m}^{-2} \text{s}^{-1}$. Variation was due to spatial variation on shelves within and among chambers. Autotrophic biomass was estimated as in ref. 22. Bacterial densities were estimated as in ref. 23.

Statistical analysis. Significant $S/F \times \text{nutrient} \times \text{light}$ effects were determined by a full-factorial, three-way analysis of variance (ANOVA) including all interaction terms.

Received 21 May; accepted 9 September 1997.

1. Folke, C., Holling, C. S. & Perrings, C. Biological diversity, ecosystems and the human scale. *Ecol. Appl.* **6**, 1018–1024 (1996).
2. May, R. M. *Stability and Complexity in Model Ecosystems* (Princeton Univ. Press, Princeton, 1974).
3. Naem, S. Species redundancy and ecosystem reliability. *Cons. Biol.* (in the press).
4. Imhoff, M. L. Mapping human impacts on the global biosphere. *BioScience* **44**, 598 (1994).
5. Heywood, V. H. *Global Biodiversity Assessment* (UNEP and Cambridge Univ. Press, Cambridge, 1995).
6. Gitay, H., Bastow, W. & Lee, W. G. Species redundancy: a redundant concept? *J. Ecol.* **84**, 121–124 (1996).
7. Walker, B. H. Biological diversity and ecological redundancy. *Cons. Biol.* **6**, 18–23 (1992).
8. Lawton, J. H. & Brown, V. K. in *Biodiversity and Ecosystem Function* (eds Schulze, E. D. & Mooney, H. A.) 255–270 (Springer, New York, 1993).
9. Tilman, D. Biodiversity: population versus ecosystem stability. *Ecology* **77**, 350–363 (1996).
10. Cowling, R. M., Mustart, P. J., Laurie, H. & Richards, M. B. Species diversity: functional diversity and functional redundancy in fynbos communities. *S. Afr. J. Sci.* **90**, 333–337 (1994).
11. Beare, M. H., Coleman, D. C., Crossley, D. A., Hendrix, P. F. & Odum, E. P. A hierarchical approach to evaluating the significance of soil biodiversity to biogeochemical cycling. *Plant Soil* **170**, 5–122 (1995).
12. Feinsinger, P., Wolfe, J. A. & Swann, L. A. Island ecology: reduced hummingbird diversity and the pollination of plants, Trinidad and Tobago, West Indies. *Ecology* **63**, 494–506 (1982).
13. Körner, C. in *Biodiversity and Ecosystem Functioning* (eds Schulze, E.-D. & Mooney, H. A.) 117–139 (Springer, Berlin, 1993).
14. Christiansen, N. L. J. in *The Ecological Basis of Conservation* (eds Pickett, S. T. A., Ostfeld, R. S., Schachak, M. & Likens, G. E.) 167–186 (Chapman and Hall, New York, 1997).
15. Baruch, Z. et al. in *Biodiversity and Savanna Ecosystem Processes* Vol. 121 (eds Solbrig, O. T., Medina, E. & Silva, J. E.) 176–194 (Springer, Berlin, 1996).
16. Harte, J. & Kinzig, A. P. Mutualism and competition between plants and decomposers: implications for nutrient allocation in ecosystems. *Am. Nat.* **141**, 829–846 (1993).
17. Drake, J. A. Community-assembly mechanics and the structure of an experimental species ensemble. *Am. Nat.* **137**, 1–26 (1991).
18. Power, M. E. et al. Challenges in the quest for keystones. *BioScience* **46**, 609–620 (1996).
19. Daily, G. C. *Nature's Services* (Island Press, Washington DC, 1997).
20. Costanza, R. et al. The value of the world's ecosystem services and natural capital. *Nature* **387**, 253–260 (1997).
21. Lawler, S. P. & Morin, P. J. Food web architecture and population dynamics in laboratory microcosms of protists. *Am. Nat.* **141**, 675–686 (1993).
22. Wetzel, R. G. & Likens, G. E. *Limnological Analyses* (Springer, Berlin, Germany, 1991).
23. Porter, K. G. & Feig, Y. S. The use of DAPI for identifying and counting aquatic microflora. *Limnol. Oceanogr.* **25**, 943–948 (1980).

Acknowledgements. We thank J. McGrady-Steed, P. Morin, B. Sterner, D. Tilman and S. Tjossem for critical evaluations of this work, and the McKnight Land-Grant Professorship and NSF for support.

Correspondence and requests for materials should be sent to S.N. (e-mail: naem001@maroon.tc.umn.edu).

A role for oestrogens in the male reproductive system

Rex A. Hess*, David Bunick*, Ki-Ho Lee*, Janice Bahr†, Julia A. Taylor‡, Kenneth S. Korach§ & Dennis B. Lubahn‡

* Department of Veterinary Biosciences, University of Illinois, 2001 S. Lincoln, Urbana, Illinois 61802, USA

† Department of Animal Sciences, University of Illinois, Urbana, Illinois 61801, USA

‡ Departments of Biochemistry and Child Health, University of Missouri, Columbia, Missouri 65211, USA

§ National Institute of Environmental Health Sciences, NIH, RTP, North Carolina 27709, USA

Oestrogen is considered to be the 'female' hormone, whereas testosterone is considered the 'male' hormone. However, both hormones are present in both sexes. Thus sexual distinctions are not qualitative differences, but rather result from quantitative divergence in hormone concentrations and differential expressions of steroid hormone receptors. In males, oestrogen is present in low concentrations in blood, but can be extraordinarily high in semen, and as high as 250 pg ml^{-1} in rete testis fluids^{1,2}, which is higher than serum oestradiol in the female³. It is well known that male reproductive tissues express oestrogen receptors^{4–7}, but the role of oestrogen in male reproduction has remained unclear.

Here we provide evidence of a physiological role for oestrogen in male reproductive organs. We show that oestrogen regulates the reabsorption of luminal fluid in the head of the epididymis. Disruption of this essential function causes sperm to enter the epididymis diluted, rather than concentrated, resulting in infertility. This finding raises further concern over the potential direct effects of environmental oestrogens on male reproduction and reported declines in human sperm counts^{8,9}.

Classic cellular responses to the hormone oestrogen are mediated through nuclear oestrogen receptors (ER), which function as ligand-dependent transcription factors. Efferent ductules of the testis are known to express high amounts of ER- α ^{10,11}, higher even than uterine tissue, and both the α and β forms of ER are present in efferent ductules and the epididymis¹⁰. These ductules form a series of small tubules that transport sperm from the testis to the epididymis¹². In humans, one third of the epididymal head consists of efferent ductules¹³. In addition to ciliated cells that stir the luminal fluid, their epithelia contain non-ciliated cells that resemble proximal tubule cells in the kidney. The non-ciliated cells have a reabsorptive function that results in the uptake of water, ions and proteins from the ductal lumen^{12,14}. Ductules in the rat reabsorb nearly 90% of the rete testis fluid, coupling water and active ion transport in an electroneutral environment, in which Na^+ and water are reabsorbed at equal rates, thereby increasing the concentration of sperm as they enter the epididymis^{15,16}. This method of concentrating sperm improves their survival and maturation during epididymal storage and ensures that a large number of sperm are released upon ejaculation, increasing the randomness of fertilization and providing genetic variation¹⁴. These data and the observation that efferent ductules contain the highest concentrations of ER in the male led us to hypothesize that oestrogen participates in the regulation of fluid reabsorption in the male reproductive tract.

To test this hypothesis, we used the ER- α gene knockout mouse (ERKO)^{17,18}. The ERKO male is infertile¹⁸, but its testes appear normal until puberty, when they begin to degenerate as early as 20–40 days of age. By 150 days, the testes are atrophic¹⁹. Sperm from the ERKO male are abnormal and sperm concentrations are significantly reduced in the epididymis¹⁹. The reproductive tract in ERKO males contains a dilated rete protruding into the testis (Fig. 1a, b). Downstream from the rete, the efferent ductules are also swollen (Fig. 1c, d), with luminal areas more than twice the size of those of wild-type males at 90 days of age. From these observations, it appears that luminal fluid is not being removed by the ductal epithelium or that there is an excess of fluid secreted by the testis, causing fluids to accumulate in seminiferous tubules, rete and efferent ductules. Epithelial cells of the wild-type (Fig. 1e) contain endocytotic vesicles and large PAS+ lysosomal granules, organelles common to cells active in the uptake of luminal fluids. However, these structures are greatly reduced or missing in the epithelium of ERKO (Fig. 1f), which is decreased in height by 45%. Based on these data, we hypothesized that an increase in testis weight would occur as testicular secretions accumulated in the lumen of the seminiferous tubule. As postulated, a transient increase in testis weight in ERKO males is seen between 32 and 81 days of age and a decrease by 185 days (Fig. 2), suggesting that long-term atrophy of testes from ERKO is caused by back-pressure of the luminal fluids^{20,21}.

We further reasoned that if the efferent ductules in ERKO mice were dysfunctional, then occlusion of the initial segment epididymis, immediately distal to the ductules, would cause fluid to accumulate more rapidly in their testes than in the wild type. Therefore, we surgically cauterized the initial segment of the epididymis, occluding the terminal end of the efferent ductules. The contralateral side served as the control. Testes were removed 48 h after surgery and the difference in weight between the occluded and sham-operated side represented the build-up of luminal fluid caused by a lack of reabsorption in the efferent ductules. After ductal occlusion, testes of ERKO weigh 30% more than testes of wild-type

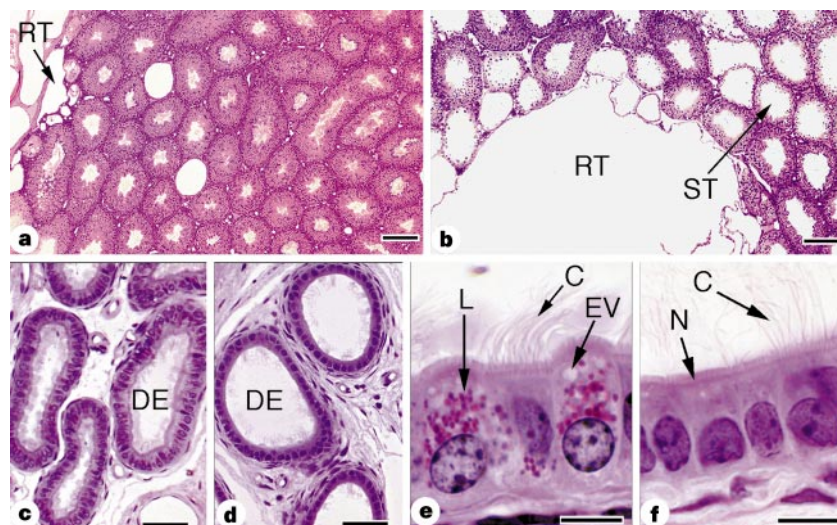


Figure 1 Seminiferous tubules, rete testis and efferent ductules in ERKO and wild-type mice. **a**, Wild-type seminiferous tubules exhibit normal spermatogenesis and a small rete testis (RT). **b**, ERKO rete testis and seminiferous tubules (ST) are dilated. Spermatogenesis appears abnormal in several tubules. **c**, Wild-type efferent ductules (DE) have narrow lumen. **d**, ERKO efferent ductules have dilated lumen. **e**, Wild-type efferent-ductule epithelium contains endocytotic vesicles (EV) and numerous PAS+ lysosomes (L). **f**, ERKO efferent-ductule epithelium is reduced in height and non-ciliated cells (N) contain fewer lysosomes and endocytotic vesicles; C, cilia. Scale bars: **a**, **b**, 100 μ m; **c**, **d**, 50 μ m; **e**, **f**, 10 μ m.

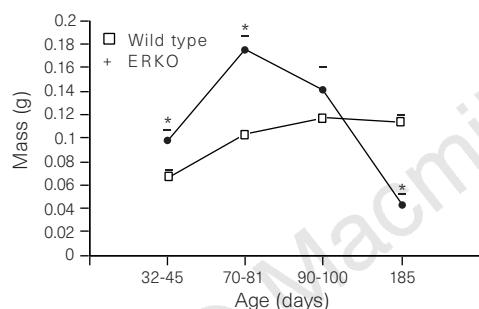


Figure 2 Testicular mass between 32 and 185 days of age (mean $\pm$ s.e.m.). ERKO mass increases at 32-45 and 70-80 days. By 185 days, the ERKO testes are atrophic. Differences were determined by the unpaired Student's *t*-test ($P < 0.05$); $N = 5-8$ for wild-type mice (WT) and $N = 7-10$ for ERKO.

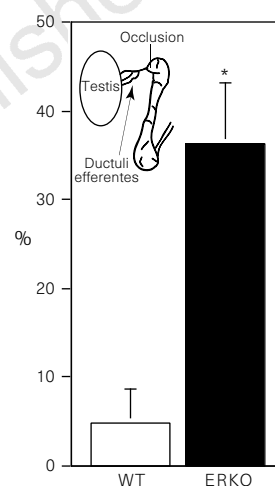


Figure 3 Change in testis mass 48h after occlusion of the initial-segment epididymis. Values shown are mean ($\pm$ s.e.m.) percentage difference between the occluded and sham-operated side. Differences between wild-type (WT) and ERKO means were determined by the unpaired Student's *t*-test ($P < 0.05$).

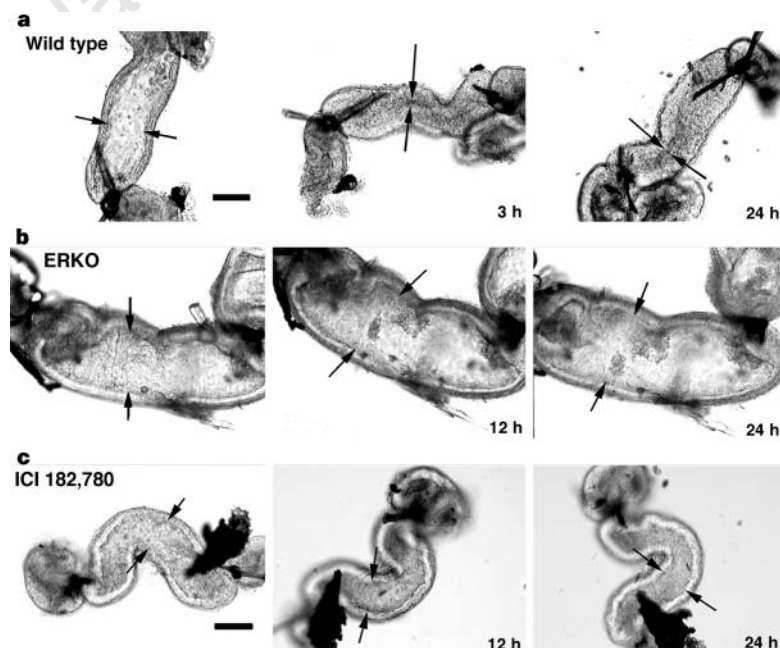


Figure 4 Ligated efferent ductules *in vitro*. **a**, Wild-type ductule lumen (between arrows) collapse by 3h. **b**, ERKO ductule lumen (between arrows) is wider than wild type and appears to increase in diameter after incubation. **c**, Wild-type ductule treated with an anti-oestrogen (ICI 182,780) initially has a normal lumen but has a slight reduction in diameter after incubation. Scale bars, 100 μ m.

mice (Fig. 3). However, these data do not remove the possibility that the increase in weight results from an abnormal rate of fluid secretion by the seminiferous epithelium in the testis of ERKO. Therefore, we also unilaterally occluded the rete testis in a group of ERKO and wild-type males and compared differences in testicular weight changes at 24 h. Instead of showing increases in fluid secretion, the testis in ERKO secretes significantly less fluid ($P < 0.05$) in 24 h than does the wild-type: $22.4 \pm 3.0\%$ increase versus $37.9 \pm 5.3\%$, respectively. In the ERKO male, efferent ductules do not appear capable of reabsorbing luminal fluids received from the testis. Thus the ERKO mouse provides evidence that oestrogens may be responsible for the regulation of fluid transport, which increases the concentration of sperm before entering the epididymis.

To assess the function of oestrogen in adult tissues without developmental influence, we treated wild-type mice for 3 days with an anti-oestrogen compound and used *in vitro* methods of analysis to determine fluid reabsorption. ICI 182,780 (ICI Pharmaceuticals) is a pure anti-oestrogen that inhibits increases in uterine weight stimulated by 17β -oestradiol and binds both ER- α and ER- β ^{22,23}. Small segments of adult efferent ductules were isolated for organ culture and the tubular ends were ligated with fine suture, preventing the inflow of culture medium. Using this *in vitro* model, we compared the function of ductules in wild-type, ERKO and ICI-treated males over a 24-h period. In ductules from wild-type mice, the epithelium is capable of rapidly removing luminal fluid, which results in the collapse of the ductule walls (Fig. 4). After 24 h of ligation, the luminal area of ductules from wild-type decreased 82% (Fig. 5). In contrast, the luminal area of ductules from ERKO increased 46%, and ductules from wild-type mice treated with ICI showed only a slight decrease of 14%. Thus efferent ductules from ERKO are incapable of reabsorbing luminal fluid, while wild-type ductules remove most of the fluid within 3 h. Blockage of ER function, by treating ductules from wild-type with an anti-oestrogen, significantly inhibits fluid reabsorption.

Thus ER- α is important both for normal growth of the male reproductive tract and for adult function of the efferent ductules. However, because wild-type ICI-treated ductules *in vitro* did not swell like the tissues from ERKO mice, the mechanisms of oestrogen action are not fully understood. For example, it is not known whether the role of ER- β is the same as that of ER- α in efferent ductules. Recent data in HeLa cells have shown that anti-oestrogens and 17β -oestradiol produce opposite effects on ER- β , depending upon the type of response element complex formed²⁴. Thus ER- β function in ductules of ERKO males could account for the *in vitro* observations. Regardless of the precise mechanisms, developmental disruption of ER- α in mice causes a total loss of net fluid reabsorption. In the human, mutational dysfunction in ER- α and P450

aromatase (the enzyme that converts androgens to oestrogens) decreases sperm counts and results in poor sperm viability^{25–27}. Our data suggest that oestrogen deficiency or oestrogen insensitivity in man might also result in the accumulation of fluid in efferent ductules and subsequent atrophy of the testis. Tamoxifen, which is commonly known to be a mixed oestrogen agonist/antagonist in different tissues, has been used to increase sperm counts in oligospermic men²⁸. In conclusion, our data describe a physiological endpoint of oestrogen action in the male reproductive system. This finding is important given recent concerns over reported declines in human sperm counts and speculation that exposure to environmental oestrogens may be a cause of this^{8,9,29}. The extensive presence of ER (both α and β) at other sites in the male reproductive system and throughout the body^{10,11,23} makes it possible that new and unexpected functions may be found for the 'female' hormone in men. □

Methods

Histology and morphometry. Testis and mid-proximal efferent ductules were fixed and processed for light microscopy³⁰. Luminal area, ductal circumference and epithelial height were taken in 5 areas per ductule cross-section in 5 ductules per mouse.

Initial-segment epididymal occlusions and rete testis ligation. Under surgical anaesthesia, initial-segment epididymidis was occluded by cauterization or 'rete' testis was ligated with 0000 suture. The contralateral testis was sham operated. The animals were allowed to recover and testes were taken at either 48 h (initial-segment occlusions) or 24 h (rete ligation) and weighed. For the initial segment occlusions, $N = 7$ for both wild-type and ERKO. For rete testis occlusions, $N = 14$ for wild-type mice and $N = 12$ for ERKO.

***In vitro* ligation of efferent ductules.** One group of wild-type mice was given single daily injections of ICI 182,780 (1 mg per kg) for 3 days before removal of the ductules. Efferent ductules were microdissected into 1.5-mm lengths and incubated for 24 h in M199 culture medium containing dihydrotestosterone (4×10^{-7} M), 17β -oestradiol (1×10^{-9} M), bovine lipoprotein (0.16 mg ml⁻¹) or all components plus ICI 182,780 (1×10^{-6} M) at 34 °C in humidified 95% air/5% CO₂. After 24 h, segments were ligated on both ends to prevent entry and exit of fluids. Digital images of the ductules were analysed at 0, 3 or 12 and 24 h after ligation (Figs 4, 5). Ductal segments damaged by microdissection or stretching were discarded. Only segments with rapid ciliary beat and clear lumens were ligated. $N = 3$ mice for each treatment group, and 3–12 ductal segments per animal were analysed. Differences between means were determined by a one-way analysis of variance ($P < 0.0001$) followed by Bonferroni multiple comparisons test ($P < 0.001$).

Received 29 July; accepted 17 September 1997.

- Ganjam, V. K. Amann, R. P. Steroid content of fluids and sperm entering and leaving the bovine epididymis, in epididymal tissue, and in accessory sex gland secretions. *Endocrinology* **99**, 1618–1630 (1976).
- Free, M. J. & Jaffe, R. A. Collection of rete testis fluid from rats without previous efferent duct ligation. *Biol. Reprod.* **20**, 269–278 (1979).
- Smith, M. S., Freeman, M. E. & Neill, J. D. The control of progesterone secretion during the estrous cycle and early pseudopregnancy in the rat: prolactin, gonadotropin and steroid levels associated with rescue of the corpus luteum of pseudopregnancy. *Endocrinology* **96**, 219–226 (1975).
- Cooke, P. S., Young, P., Hess, R. A. & Cunha, G. R. Estrogen receptor expression in developing epididymis, efferent ductules, and other male reproductive organs. *Endocrinology* **128**, 2874–2879 (1991).
- Greco, T., Duello, T. & Gorski, J. Estrogen receptors; estradiol; and diethylstilbestrol in early development: the mouse as a model for the study of estrogen receptors and estrogen sensitivity in embryonic development of male and female reproductive tracts. *Endocr. Rev.* **14**, 59–71 (1993).
- Schleicher, G., Drews, U., Stumpf, W. E. & Sar, M. Differential distribution of dihydrotestosterone and estradiol binding sites in the epididymis of the mouse. An autoradiographic study. *Histochemistry* **81**, 139–147 (1984).
- West, N. & Brenner, R. Estrogen receptor in the ductuli efferentes epididymis and testis of rhesus and cynomolgus macaques. *Biol. Reprod.* **42**, 533–538 (1990).
- Sharpe, R. M. & Skakkebaek, N. E. Are oestrogens involved in falling sperm counts and disorders of the male reproductive tract? *Lancet* **341**, 1392–1395 (1993).
- Auger, J., Kunstmann, J. M., Czyglik, F. & Jouannet, P. Decline in semen quality among fertile men in Paris during the past 20 years. *N. Engl. J. Med.* **332**, 281–285 (1995).
- Hess, R. A. *et al.* Estrogen receptor (α & β) expression in the excurrent ducts of the adult male rat reproductive tract. *J. Androl.* **18**, 602–611 (1997).
- Fisher, J. S. *et al.* Immunolocalisation of estrogen receptor- α within the testis and excurrent ducts of the rat and marmoset monkey from perinatal life to adulthood. *J. Endocrinol.* **153**, 485–495 (1997).
- Ilio, K. & Hess, R. Structure and function of the ductuli efferentes: A review. *Microsc. Res. Tech.* **29**, 432–467 (1994).
- Yeung, C. H., Cooper, T. G., Bergmann, M. & Schulze, H. Organization of tubules in the human caput epididymidis and the ultrastructure of their epithelia. *Am. J. Anat.* **191**, 261–279 (1991).

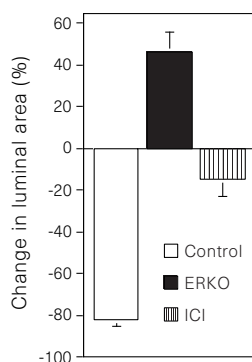


Figure 5 The effects of *in vitro* ligation on changes in luminal area of efferent ductule segments 24 h after ligation. Mean ($\pm$ s.e.m.) percentage difference between 0 h and 24 h in luminal areas for isolated ductules from wild-type controls, ERKO and ICI-treated mice.

14. Robaire, B. & Hermo, L. in *The Physiology of Reproduction* (eds Knobil, E. & Neill, J.) 999–1080 (Raven, New York, 1988).
15. Chan, H. C., Zhou, W. L., Fu, W. O., Ko, W. H. & Wong, P. Y. Different regulatory pathways involved in ATP-stimulated chloride secretion in rat epididymal epithelium. *J. Cell. Physiol.* **164**, 271–276 (1995).
16. Clulow, J., Jones, R. & Hansen, L. Micropuncture and cannulation studies of fluid composition and transport in the ductuli efferentes testis of the rat: comparisons with the homologous metanephric proximal tubule. *Exp. Physiol.* **79**, 915–928 (1994).
17. Lubahn, D. B. *et al.* Alteration of reproductive function but not prenatal sexual development after insertional disruption of the mouse estrogen receptor gene. *Proc. Natl Acad. Sci. USA* **90**, 11162–11166 (1993).
18. Korach, K. S. *et al.* Estrogen receptor gene disruption: molecular characterization and experimental and clinical phenotypes. *Recent Prog. Horm. Res.* **51**, 159–186 (1996).
19. Eddy, E. M. *et al.* Targeted disruption of the estrogen receptor gene in male mice causes alteration of spermatogenesis and infertility. *Endocrinology* **137**, 4796–4805 (1996).
20. Smith, G. The effects of ligation and the vasa efferentia and vasectomy on testicular function in the adult rat. *Endocrinology* **23**, 385–399 (1962).
21. Hess, R., Moore, B., Forrer, J., Linder, R. & Abuel-Atta, A. The fungicide benomyl [methyl 1-butylcarbamoyl-2-benzimidazole carbamate] causes testicular dysfunction by inducing the sloughing of germ cells and occlusion of efferent ductules. *Fundam. Appl. Toxicol.* **17**, 733–745 (1991).
22. Wakeling, A., Dukes, M. & Bowler, J. A potent specific pure antiestrogen with clinical potential. *Cancer Res.* **51**, 3867–3873 (1991).
23. Kuiper, G. G. *et al.* Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors alpha and beta. *Endocrinology* **138**, 863–870 (1997).
24. Paech, K. *et al.* Differential ligand activation of estrogen receptors ER α and ER β at AP1 sites. *Science* **277**, 1508–1510 (1997).
25. Carani, C. *et al.* Effect of testosterone and estradiol in a man with aromatase deficiency. *N. Engl. J. Med.* **337**, 91–95 (1997).
26. Morishima, A., Grumbach, M. M., Simpson, E. R., Fisher, C. & Qin, K. Aromatase deficiency in male and female siblings caused by a novel mutation and the physiological role of estrogens. *J. Clin. Endocrinol. Metab.* **80**, 3689–3698 (1995).
27. Smith, E. P. *et al.* Estrogen resistance caused by a mutation in the estrogen-receptor gene in a man. *N. Engl. J. Med.* **331**, 1056–1061 (1994). Erratum. *N. Engl. J. Med.* **332**, 131 (1995).
28. Krause, W., Holland-Moritz, H. & Schramm, P. Treatment of idiopathic oligozoospermia with tamoxifen—a randomized controlled study. *Int. J. Androl.* **15**, 14–18 (1992).
29. Sharpe, R. M. Declining sperm counts in men—is there an endocrine cause? *J. Endocrinol.* **136**, 357–360 (1993).
30. Hess, R. A. & Moore, B. J. in *Methods in Reproductive Toxicology* (eds Chapin, R. E. & Heindel, J. J.) 52–85 (Academic, San Diego, 1993).

Acknowledgements. We thank E. Jassim and C. Finnigan-Bunick for technical assistance, P. Cooke, V. K. Ganjam and D. J. Miller for reviews of the manuscript and A. Wakeling and Zeneca Pharmaceuticals for providing ICI 182, 780.

Correspondence and requests for materials should be addressed to R.A.H. (e-mail: r-hess@uiuc.edu).

Radial optic flow induces vergence eye movements with ultra-short latencies

C. Busettini*, G. S. Masson* & F. A. Miles

Laboratory of Sensorimotor Research, National Eye Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

An observer moving forwards through the environment experiences a radial pattern of image motion on each retina. Such patterns of optic flow are a potential source of information about the observer's rate of progress¹, direction of heading² and time to reach objects that lie ahead³. As the viewing distance changes there must be changes in the vergence angle between the two eyes so that both foveas remain aligned on the object of interest in the scene ahead. Here we show that radial optic flow can elicit appropriately directed (horizontal) vergence eye movements with ultra-short latencies (roughly 80 ms) in human subjects. Centrifugal flow, signalling forwards motion, increases the vergence angle, whereas centripetal flow decreases the vergence angle. These vergence eye movements are still evident when the observer's view of the flow pattern is restricted to the temporal hemifield of one eye, indicating that these responses do not result from anisotropies in motion processing but from a mechanism that senses the radial pattern of flow. We hypothesize that flow-induced vergence is but one of a family of rapid ocular reflexes, mediated by the medial superior temporal cortex, compensating for translational disturbance of the observer.

* Present addresses: Department of Physiological Optics, University of Alabama, Birmingham, Alabama 35205, USA (C.B.); Centre de Recherche en Neurosciences Cognitives, Centre National de la Recherche Scientifique, 13402 Marseille, France (G.S.M.).

The centrifugal pattern of optic flow that we experience as we move forwards—a constant stream of images moving away from a central focus of expansion (Fig. 1a)—has been the subject of many psychophysical, physiological and computational studies (see ref. 4 for a review). Optic flow is a rich source of information about the three-dimensional layout of the environment and the observer's movements within it⁵. To maintain binocular alignment on objects of interest that lie ahead, the observer must continuously adjust the vergence angle between the eyes. The primary visual cue used to control convergence is binocular disparity⁶, though changes in object size^{7,8} as well as static perspective and overlay cues⁹ can also be used. It occurred to us that optic flow might also make a contribution as it charts the observer's rate of forward progress.

Our experimental subjects faced a vertical screen onto which large random-dot patterns were back-projected. Steps of radial flow, achieved by switching between two projected images signalling radial displacement and a size change, were applied 50 ms after the subject's eyes fixated the centre of the screen. The flow patterns elicited pure horizontal vergence (negligible vertical vergence) at latencies of roughly 80 ms in the four subjects tested. As expected from the geometry, centrifugal flow (signalling approach) resulted in increased convergence whereas centripetal flow (signalling retreat) resulted in decreased convergence. Sample vergence velocity profiles, obtained with centrifugal steps of various amplitudes (expressed in terms of the percentage change in apparent viewing distance), are shown for one subject in Fig. 1b. Initial responses were quantified by measuring the change in vergence over a 33-ms period commencing a fixed time (95 ms) after the onset of the stimulus. This meant that our measures were restricted to the initial (open-loop) vergence responses that were generated by the radial flow before it had been affected by eye-movement feedback. These measures are plotted against the percentage change in apparent viewing distance in Fig. 1c, for both centrifugal and centripetal flow. For a given step size, the increases in convergence with centrifugal flow were always larger than the decreases with centripetal flow (compare with ref. 10). Optimal steps changed the apparent viewing distance by 2–4%.

Vergence velocity profiles were always transient, rising to a peak in 30–50 ms and then declining. In our artificial viewing situation, the vergence responses to radial flow would result in binocular disparity, which would in turn tend to generate vergence in the opposing direction. However, disparity-induced vergence has been shown to have a latency approximating that of the radial-flow vergence¹¹, indicating that its earliest effects could not occur until about 80 ms after response onset. The transient nature of the radial-flow vergence cannot therefore be due to the 'closing of the disparity-vergence loop'. This conclusion is also supported by the finding that initial vergence responses were still transient with monocular viewing (discussed later).

When an observer moves forwards, retinal images not only undergo centrifugal motion but also get larger (Fig. 2a), as was the case in the above experiments. Changes in the size of a single small foveated target have been shown to elicit weak vergence eye movements in humans, but at latencies estimated to be in excess of 200 ms (refs. 7, 8). In our case, using large multi-element patterns and shorter time intervals, the size changes were irrelevant, and eliminating the size cue (as in Fig. 2b) had little effect on the vergence responses. Vergence responses to radial flow alone were almost the same as those to radial flow plus size change, and pure changes in size (as in Fig. 2c) failed to generate significant vergence (Fig. 2d).

We propose that these vergence eye movements represent a binocular response to the radial pattern of flow. However, an alternative explanation might be that tracking is essentially monocular, each eye tracking only the motion that it sees and with a strong preference for motion in the nasal hemifields. In this case, with centrifugal flow there would be a net motion vector towards

the nose in the nasal hemifields and each eye would independently track that motion, the result being increased convergence. With centripetal flow there would be a net motion vector away from the nose in the nasal hemifields and individual eye tracking would result in decreased convergence. Masking different parts of the flow fields seen by each eye shows that this is not the correct explanation (Fig. 3). When one eye was covered, both eyes still converged, albeit more weakly. Further, the convergence persisted when both nasal hemifields were masked off. Here each eye moved in the opposite direction to the net motion vector that it saw. For example, the right eye saw net motion to the right and yet moved to the left. The convergence even persisted when the subject's view was restricted to only one temporal hemifield. The latter result is particularly revealing about the sensing mechanism that decodes the flow pattern. Despite there being a clear net motion vector to the right, the system still responded correctly, as though compensating for forwards (rather than leftwards) motion, hence the convergent (rather than conjugate rightwards) response. Presumably, the

(opposed) vertical components of the flow pattern are critical here for the system to draw this distinction between radial flow (signalling forwards and backwards motion) and planar flow (signalling side-to-side motion). We obtained very similar data when the temporal hemifields were masked off, though responses were sometimes a little better. We conclude that the vergence responses to radial flow result from a true parsing of the flow pattern and not from anisotropies in motion detection.

Two other types of eye movement can be elicited with similar ultra-short latencies by large-field visual stimuli. There are indications that they too help to compensate specifically for translational disturbances of the observer¹². The first employs conjugate eye movements (often termed 'ocular following') to track binocular motion in the plane of fixation as though compensating for lateral or vertical disturbances in the frontoparallel plane¹³⁻¹⁵, and the second employs vergence eye movements to eliminate binocular misalignments using binocular disparity cues ('disparity vergence') and so, like the radial-flow vergence described here, deals with

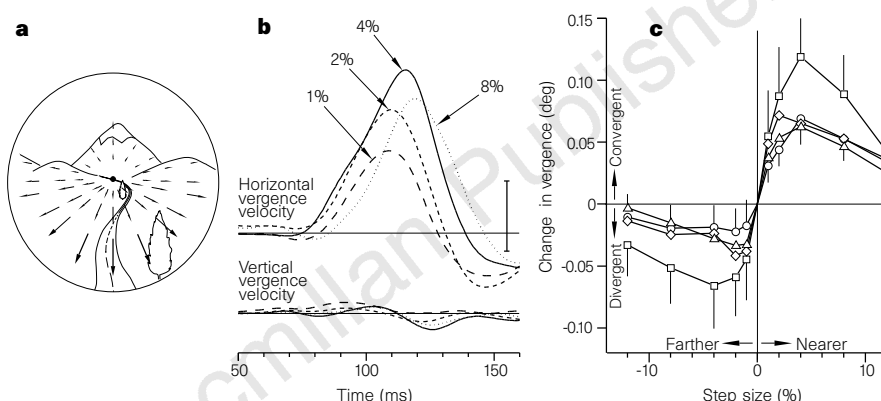


Figure 1 Initial vergence responses to looming steps (radial optic flow plus size change). **a**, Centrifugal pattern of optic flow experienced by an observer moving forwards (towards the black dot at the foot of the mountain). The amount of convergence required to maintain binocular alignment during forward movement is inversely related to the viewing distance, so the vergence required with the distant viewing depicted here would be minimal. Our experiments were all conducted with near viewing. **b**, Mean horizontal and vertical vergence velocity profiles of subject GSM in response to looming steps simulating sudden

decreases in the viewing distance of 1–8% at time zero. Upwards deflections indicate increased convergence and left sursumvergence. Calibration bar, 2° s^{-1} . **c**, Plot of mean ($\pm$ s.d.) changes in vergence measured over time period 95–128 ms (from onset of stimulus) against the simulated percentage change in viewing distance for four human subjects: square, subject GSM; circle, FAM; triangle, MAB; diamond, RJK. Each trace in **b** and each point in **c** is based on at least 80 responses.

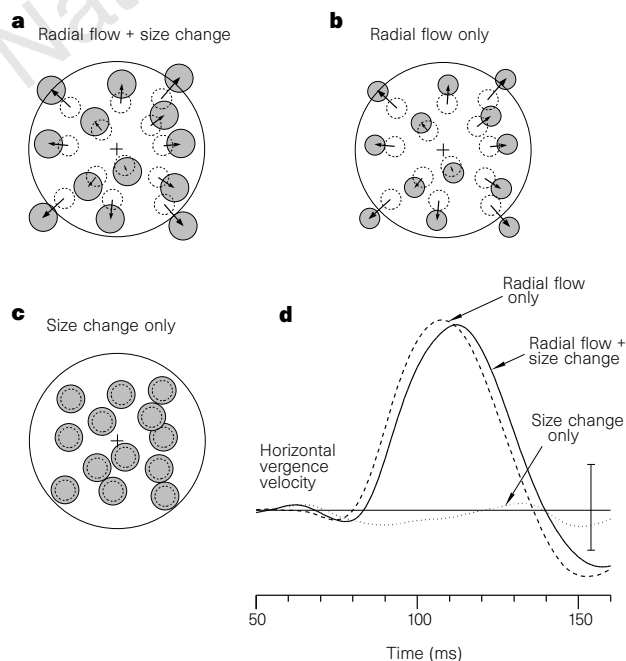


Figure 2 The adequate stimulus for the short-latency vergence associated with looming steps is radial flow and the size changes are irrelevant. **a–c**, Stimuli used (not drawn to scale) to test the dependence of vergence on radial flow and size change. **a**, Radial flow plus size change (each dot shifts peripherally and enlarges). **b**, Radial flow alone (each dot shifts peripherally but stays the same size). **c**, Size change alone (each dot enlarges in place). **d**, Mean vergence-velocity profiles of subject FAM in response to the three types of stimuli applied at time zero (simulating 4% reduction in the viewing distance). Upwards deflections indicate increased convergence. Calibration bar, 2° s^{-1} . All profiles are means of at least 138 responses.

front-to-back disturbances^{11,16}. Radial-flow vergence shares an additional property with these other ocular responses—transient post-saccadic enhancement—whereby stimuli applied (as here) in the immediate wake of a saccadic eye movement are much more effective than the same stimuli applied several hundred milliseconds later^{11,13}. In all cases, the post-saccadic enhancement is at least partly visual in origin, resulting from the reafferent visual stimulation created by the saccade as it sweeps the image of the pattern across the retina: all show a similar (though often somewhat weaker) transient enhancement in the wake of a saccade-like shift of the large-field pattern. Appropriate large-field stimuli have also been shown to elicit ocular following¹⁷ and disparity-vergence^{16,18} responses in monkeys with properties often strikingly similar to those of humans.

Visual motion is processed mainly in the so-called dorsal stream of cortex¹⁹. In monkeys, the medial superior temporal (MST) region of the cortex is the earliest stage at which global optic flow is

encoded at the level of single neurons^{20–25}. Of particular interest here are MST neurons that are selectively sensitive to radial optic flow and do not respond to component size change²². We suggest that all three types of eye movements elicited at ultra-short latencies are mediated by MST or its human homologue²⁶. In the case of ocular following, support comes from single-unit recordings and chemical lesions in monkeys^{27,28}. Putative ocular following neurons in MST respond to large-field motion in the frontoparallel plane with strong directional selectivity, being excited by motion in one direction and inhibited by motion in the opposite or orthogonal directions, thereby avoiding activation by radial flow²³. Such orthogonal inhibition might explain how the system could correctly attribute the flow experienced by our subjects to forwards, rather than to lateral, motion when exposure to the flow pattern was limited to one temporal hemifield. Recent recordings in MST have revealed neurons that discharge in relation to the short-latency disparity-vergence responses with disparity tuning curves that exactly match the behaviour, unlike earlier stages in the cortical pathway to MST²⁹. In summary, we suggest that the flow-induced vergence described here is one of a family of visually driven eye movements in monkeys and humans that are elicited with ultra-short latencies, mediated by MST, and help to compensate specifically for translational disturbances of the observer. □

Methods

Most techniques have been described previously^{13,18}. The positions of both eyes were recorded using the electromagnetic search-coil technique³⁰. Subjects faced a vertical screen (viewing distance 33 cm; subtense, 80° × 80°) onto which random-dot patterns (dot size, 1°) that filled the screen were back-projected from one of two slide projectors. The projected images were positioned with x/y mirror galvanometers and their presentation controlled by electromechanical shutters. Trials started with one pattern in place, and a target spot produced by a light-emitting diode (LED) was projected onto the screen, 10° to the right of centre. The subject was required to fixate this spot for a randomized time interval, after which the spot disappeared and a second appeared at the centre of the screen. Subjects were required to make a saccadic eye movement to this new spot, at which time it was switched off. With gaze now directed at the centre of the screen, after a post-saccadic delay of 50 ms the initial pattern was replaced with a new image (simulating the same pattern viewed from a different distance) produced by the second projector. After 200 ms the screen was blanked, ending the trial. Successive slides were carefully aligned so that the apparent focus of expansion (or contraction) was exactly at the centre of the screen and so imaged in the fovea. Applying optic flow stimuli after centring saccades took advantage of post-saccadic enhancement. For the hemifield experiments, four additional projectors were used with orthogonal polarizing filters in the projection paths to ensure that each eye saw only one hemifield¹⁸. For monocular tests, one eye was occluded with a shutter that closed during the centering saccade.

Received 31 July; accepted 2 September 1997.

- Prokop, T., Schubert, M. & Berger, W. Visual influence on human locomotion: Modulation to changes in optic flow. *Exp. Brain Res.* **114**, 63–70 (1997).
- Warren, W. H. & Hannon, D. J. Direction of self-motion is perceived from optical flow. *Nature* **336**, 162–163 (1988).
- Lee, D. N. A theory of visual control of braking based on information about time-to-collision. *Perception* **5**, 437–459 (1976).
- Cornilleau-Péres, V. & Gielen, C. C. A. M. Interactions between self-motion and depth perception in the processing of optic flow. *Trends Neurosci.* **19**, 196–202 (1996).
- Lee, D. N. The optic flow field: the foundation of vision. *Phil. Trans. R. Soc. Lond. B* **290**, 169–179 (1980).
- Collewijn, H. & Erkelens, C. J. in *Eye Movements and their Role in Visual and Cognitive Processes* (ed. Kowler, E.) 213–261 (Elsevier, Amsterdam, 1990).
- Erkelens, C. J. & Regan, D. Human ocular vergence movements induced by changing size and disparity. *J. Physiol. (Lond)* **379**, 145–169 (1986).
- Cohen, G. A. & Lisberger, S. G. Motion disparity and looming cues form hierarchical inputs for smooth pursuit of targets moving in 3 dimensions. *Soc. Neurosci. Abstr.* **22**, 964 (1996).
- Enright, J. T. Art and the oculomotor system: perspective illustrations evoke vergence changes. *Perception* **16**, 731–746 (1987).
- Takeuchi, T. Visual search of expansion and contraction. *Vision Res.* **37**, 2083–2090 (1997).
- Bussetini, C., Miles, F. A. & Krauzlis, R. J. Short-latency disparity vergence responses in humans. *Soc. Neurosci. Abstr.* **20**, 1403 (1994).
- Miles, F. A. in *Eye Movement Research: Mechanisms, Processes and Applications* (eds Findlay, J. M., Kentridge, W. & Walker, R.) 47–62 (Elsevier, Amsterdam, 1995).
- Gellman, R. S., Carl, J. R. & Miles, F. A. Short latency ocular-following responses in man. *Vis. Neurosci.* **5**, 107–122 (1990).

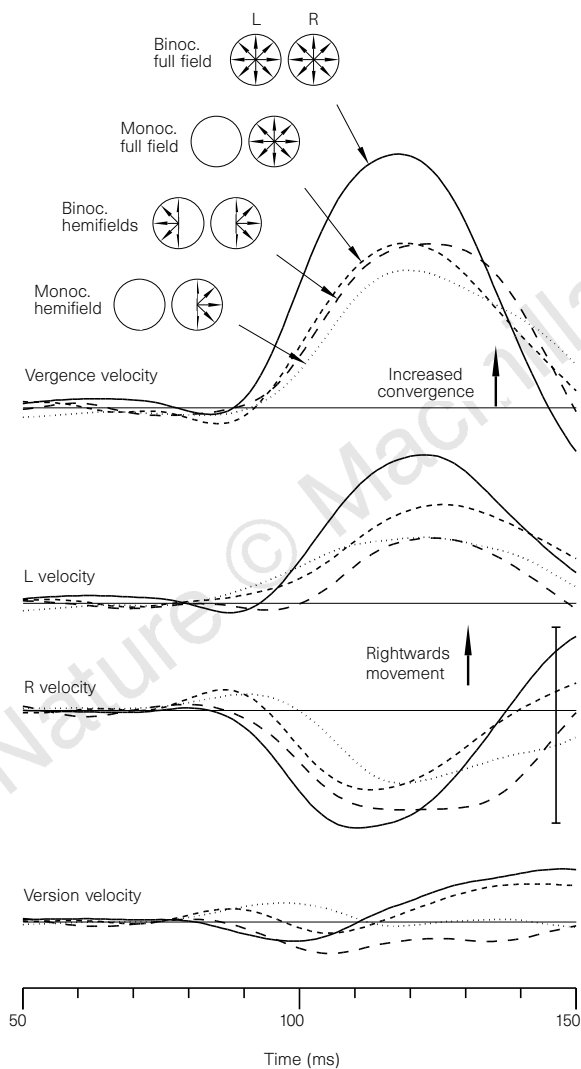


Figure 3 Effect of masking off various parts of the binocular visual fields on initial ocular responses of subject FAM to centrifugal optic flow. Insets indicate the extent of the masks: no mask ('binoc. full field'), left eye masked ('monoc. full field'), both nasal hemifields masked ('binoc. hemifields') and all but one temporal hemifield masked ('monoc. hemifield'). In addition to the usual vergence velocity profiles (the difference between the velocity of the two eyes), the velocity profiles for each eye are shown, as well as the version velocity (the average velocity of the two eyes). All stimuli, simulating 4% reduction in viewing distance, occurred at time zero. Calibration bar, 2° s⁻¹. All profiles are means of at least 110 responses.

14. Busetini, C., Miles, F. A., Schwarz, U. & Carl, J. R. Human ocular responses to translation of the observer and of the scene: dependence on viewing distance. *Exp. Brain Res.* **100**, 484–494 (1994).
15. Busetini, C., Masson, G. S. & Miles, F. A. A role for stereoscopic depth cues in the rapid visual stabilization of the eyes. *Nature* **380**, 342–345 (1996).
16. Masson, G. S., Busetini, C. & Miles, F. A. Vergence eye movements in response to binocular disparity without the perception of depth. *Nature* **389**, 283–286 (1997).
17. Miles, F. A., Kawano, K. & Optican, L. M. Short-latency ocular following responses of monkey. I. Dependence on temporospatial properties of the visual input. *J. Neurophysiol.* **56**, 1321–1354 (1986).
18. Busetini, C., Miles, F. A. & Krauzlis, R. J. Short-latency disparity vergence responses and their dependence on a prior saccadic eye movement. *J. Neurophysiol.* **75**, 1392–1410 (1996).
19. Ungerleider, L. G. & Mishkin, M. in *Analysis of Visual Behavior* (eds Ingle, D. J., Goodale, M. A. & Mansfield, R. J. W.) 549–586 (MIT Press, Cambridge, 1982).
20. Saito, H. et al. Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. *J. Neurosci.* **6**, 145–157 (1986).
21. Tanaka, K. & Saito, H. Analysis of motion of the visual field by direction, expansion/contraction, and rotation cells clustered in the dorsal part of the medial superior temporal area of the macaque monkey. *J. Neurophysiol.* **62**, 626–641 (1989).
22. Tanaka, K., Fukada, Y. & Saito, H. Underlying mechanisms of the response specificity of expansion/contraction and rotation cells in the dorsal part of the medial superior temporal area of the macaque monkey. *J. Neurophysiol.* **62**, 642–656 (1989).
23. Duffy, C. J. & Wurtz, R. H. Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli. *J. Neurophysiol.* **65**, 1329–1345 (1991).
24. Duffy, C. J. & Wurtz, R. H. Response of monkey MST neurons to optic flow stimuli with shifted centers of motion. *J. Neurosci.* **15**, 5192–5208 (1995).
25. Lappe, M., Bremmer, F., Pökel, M., Thiele, A. & Hoffmann, K.-P. Optic flow processing in monkey STS: A theoretical and experimental approach. *J. Neurosci.* **16**, 6265–6285 (1996).
26. de Jong, B. M., Shipp, S., Skidmore, B., Frackowiak, R. S. J. & Zeki, S. The cerebral activity related to the visual perception of forward motion in depth. *Brain* **117**, 1039–1054 (1994).
27. Kawano, K., Shidara, M., Watanabe, Y. & Yamane, S. Neural activity in cortical area MST of alert monkey during ocular following responses. *J. Neurophysiol.* **71**, 2305–2324 (1994).
28. Kawano, K., Inoue, Y., Takemura, A., Kitama, T. & Miles, F. A. in *Parietal Lobe Contributions to Orientation in 3D Space* (eds Thier, P. & Karnath, H.-O.) 185–199 (Springer, Heidelberg, 1997).
29. Takemura, A., Inoue, Y., Kawano, K. & Miles, F. A. Short-latency discharges in medial superior temporal area of alert monkeys to sudden changes in the horizontal disparity. *Soc. Neurosci. Abstr.* **23**, 1557 (1997).
30. Fuchs, A. F. & Robinson, D. A. A method for measuring horizontal and vertical eye movement chronically in the monkey. *J. Appl. Physiol.* **21**, 1068–1070 (1966).

Acknowledgements. G.S.M. was supported by La Fondation pour la Recherche Médicale (France).

Correspondence and requests for materials should be addressed to F.A.M. (e-mail: fam@lsr.nci.nih.gov).

Defective limbic system in mice lacking the *tailless* gene

A. P. Monaghan*, D. Bock*, P. Gass*, A. Schwäger*,
D. P. Wolfer†, H.-P. Lipp† & G. Schütz*

* Division of Molecular Biology of the Cell I, German Cancer Research Centre, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

† Institute of Anatomy, University of Zürich-Irchel, 8057 Zürich, Switzerland

The gene *tailless* is a member of the superfamily of genes that encode transcription factors of the ligand-activated nuclear receptor type, and is expressed in the invertebrate and vertebrate brain^{1–4}. In mice, its transcripts are restricted to the periventricular zone of the forebrain⁴, the site of origin of neurons and glia. Here we use homologous recombination to generate mice that lack a functional *tailless* protein. Homozygous mutant mice are viable at birth, indicating that *tailless* is not required for prenatal survival; however, adult mutant mice show a reduction in the size of rhinencephalic and limbic structures, including the olfactory, infrarhinal and entorhinal cortex, amygdala and dentate gyrus. Both male and female mice are more aggressive than usual and females lack normal maternal instincts. These animals therefore enable a molecular approach to be taken towards understanding the genetic architecture and morphogenesis of the forebrain.

Nuclear receptors are ligand-activated transcription factors, many of which transduce extracellular signals to the nucleus; this family includes receptors for steroids, retinoids, thyroid hormone and vitamin D (ref. 5). The *tailless* gene encodes a divergent member of this superfamily and belongs to the orphan receptor sub-branch because its ligand (if it exists) has not yet been identified¹. The *Drosophila* gene (*tll*) and its chicken homologue bind a unique site (AAGTA) not recognized by other members of the nuclear-receptor superfamily^{2,3}. In the mouse, transcription of the *tailless* (*tlx*) gene is initially detected on day 8.25 post coitum (p.c.) (ref. 4). During

development its transcripts are restricted to the telencephalon, diencephalon, eye and nasal placode. To define the contribution of the *tlx* gene to mouse brain development and function, we have generated a targeted disruption of the gene by homologous recombination. The targeting strategy of the *tlx* locus is indicated (Fig. 1a). We removed exons 2 and 3, which encode the two zinc-fingers (Fig. 1a). Homozygous animals were generated at the expected ratio by heterozygote crossing (Fig. 1b). Northern blot analysis revealed that the 3-kilobase (kb) *tlx* transcript was missing from mutant brains (results not shown), and RNase protection analysis confirmed that the sequences encoding the zinc-fingers were absent (Fig. 1c); some residual 3' RNA did exist (exon 5; 20% of wild type; results not shown), however, probably as a result of inefficient termination signals in the neomycin-resistance gene or of alternative splicing around the insertion—either would cause a frameshift mutation. We concluded that the *tlx* gene had been effectively disrupted.

At birth, *tlx*^{−/−} mice are indistinguishable from their littermates, but over their first three weeks they fail to gain weight normally, unlike mutant animals that are reared alone or hand-fed (Fig. 2a). Examination of ten brains from eight-week-old animals showed that all had a visible reduction in the size of their posterior cerebral hemispheres (Fig. 2c). Histological analysis revealed that the cytoarchitecture of the neocortex above the rhinal sulcus appeared to be normal; however, the so-called rhinencephalon, including the piriform cortex, the islands of Calleja, the corticomedial amygdala, and the entorhinal cortex were markedly reduced (Figs 3, 4). Sections through the hippocampus revealed a much smaller dentate gyrus (Figs 3f, h, 4d, 5f) associated with a reduction and an abnormal lamination of the mossy fibre projection (Fig. 4f). The remainder of the hippocampus, the subfields CA3 and CA1, and the subiculum appeared normal. Pre- and parasubiculum were reduced or missing (Fig. 4f). The fornix and hippocampal commissure were abnormally shaped (not shown). The anterior commissure was also reduced, in particular its anterior wings connecting the olfactory cortices (Fig. 3d). Additional anatomical defects outside the limbic system were not found in the mice examined.

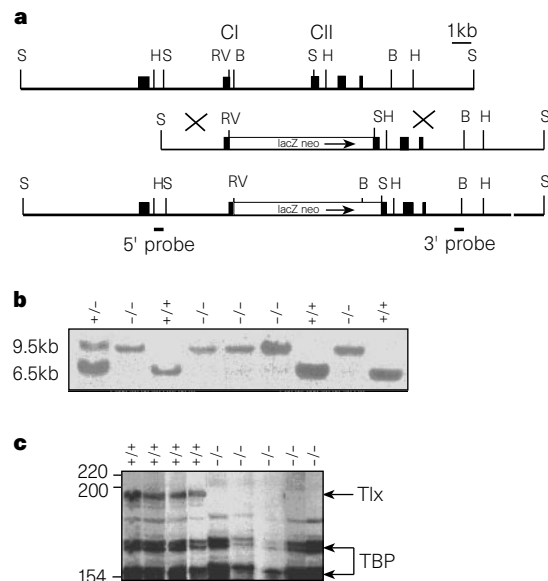


Figure 1 Targeted disruption of the *tlx* gene. **a**, A restriction map of the first five exons surrounding the two zinc-fingers (CI and CII) is shown, together with that of the properly disrupted locus. The location of probes 5' and 3' are indicated. **b**, Southern blot analysis of *Hind*III-digested (H) DNA from nine mice hybridized with the 5' probe (+/+ 6.5 kb and −/− 9.5 kb). **c**, *Bam*HI; RV, *Eco*RV; and S, *Sal*I. **c**, RNase protection analysis of 20 µg of total brain RNA hybridized with a 180-bp probe covering the zinc-fingers, which is absent in mutant animals. TATA-box-binding protein (TBP) serves as a loading control²⁰.

The distribution of specific neuronal populations in the cortex and hippocampus was studied immunohistologically (Fig. 4). The staining pattern of Krox-24 in the CA1 region was unaffected and fewer calretinin-positive neurons were detected (Fig. 4b, d). The dramatic reduction in size of the dentate gyrus is evident from the pattern of calretinin and Timm staining (Fig. 4d, f), which reveal a dramatically reduced number of granule cells which connect abnormally to CA3. The normal termination of mossy fibres on

apical dendrites above the pyramidal cell layer has shifted to the basal dendrites, producing a massive and abnormal infrapyramidal mossy fibre projection (Fig. 4f). The six-layered cytoarchitecture of the cortex appears intact. However, there is a significant decrease in the number of somatostatin- (not shown) and calretinin-positive neurons in the cortex (Fig. 4h). In contrast, the distribution of neurons containing parvalbumin, the transcription factor Krox-24, calbindin, and nitric oxide synthase appears normal (not shown).

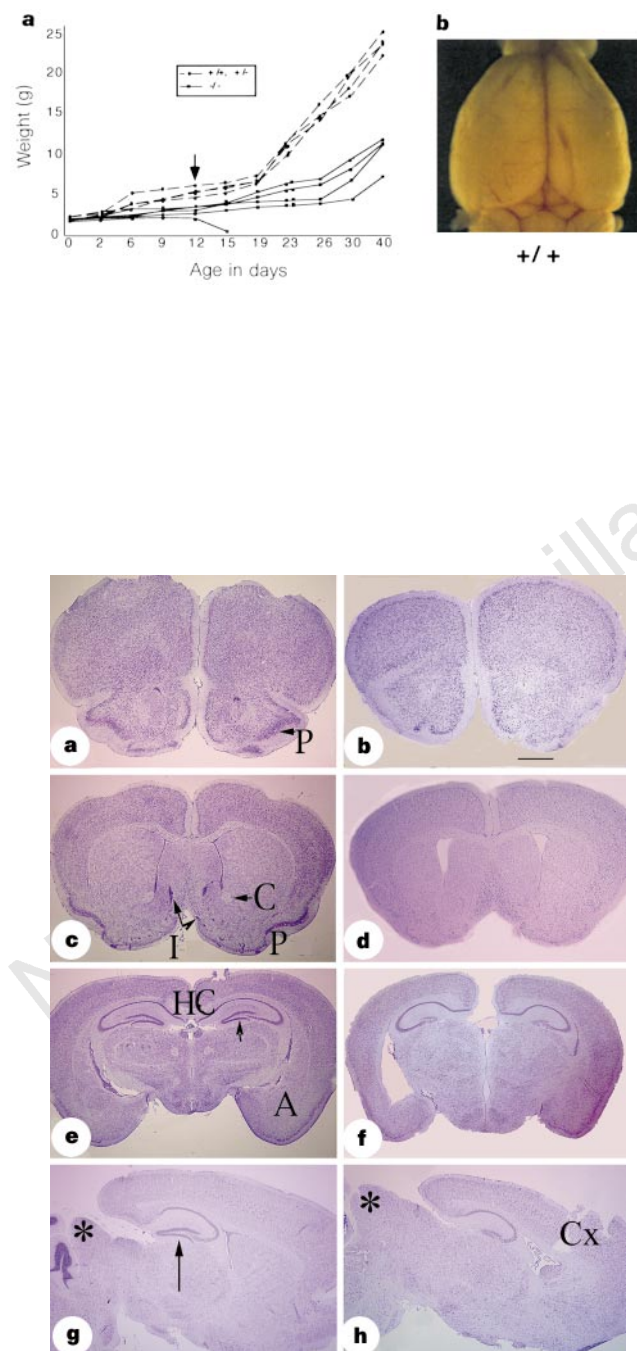


Figure 3 Reduced size of hemispheres of *tailless* mutants is due to defective rhinencephalic structures. (+/+ **a, c, e, g**; +/- **b, d, f, h**). **a-d**, Nissl-stained coronal sections showing reduction of the piriform cortex (P) and absence of the islands of Calleja (I). The anterior wings of the anterior commissure (C) are missing. **e-f**, Coronal, and **g-h**, sagittal sections through the hippocampus (HC), showing a reduction in the amygdala (A), dentate gyrus (arrows) and cortex (Cx). The colliculi (star) protrude through the cortex in mutant brains. Scale bar: **a**, 920 μ m; **b**, 610 μ m; **c, d**, 980 μ m; **e-h**, 1.2 mm.

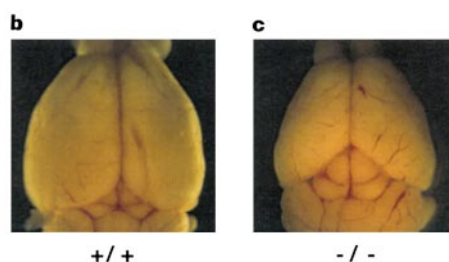


Figure 2 Growth of *tailless* mutant mice is retarded and they have smaller cerebral hemispheres. **a**, Weight curves of nine littermates after birth. The arrow indicates the age at which animals started to be hand-fed. Brains from **b**, 8-week-old wild-type, and **c**, homozygous mutant mice, showing the reduction in size of the cerebral hemisphere.

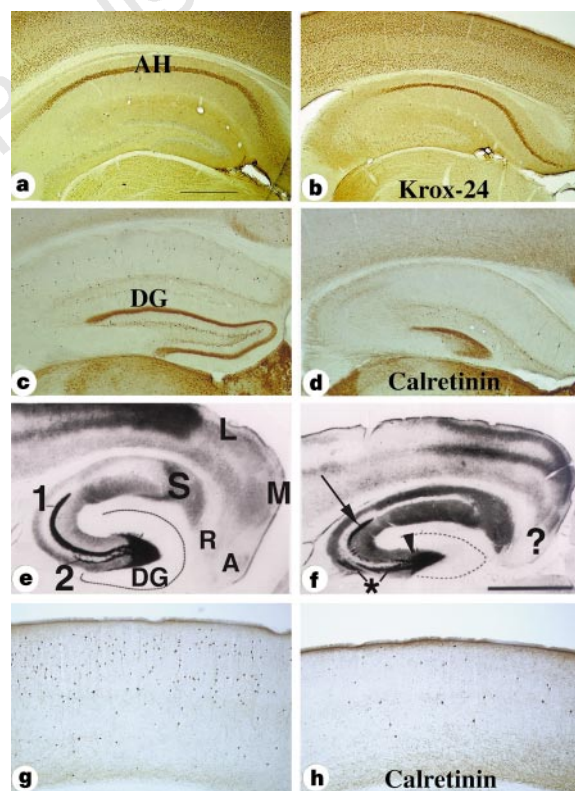


Figure 4 Reduction of neuronal subpopulations in *tlx* mutant mice. The hippocampus of 8-week-old littermates (+/+ **a, c, e**; +/- **b, d, f**) coronally sectioned and stained with antibodies to Krox-24 (**a, b**) and calretinin (**c, d**). Ammons horn (AH) appears normal, but the dentate gyrus (DG) is smaller. **e, f**, Timm-stained horizontal sections showing deficient entorhinal input to an intact hippocampal formation. In the wild-type animal, the lateral (L) and medial entorhinal cortex (M) are clearly delineated, as are the para-subiculum (A), pre-subiculum (R) and the subiculum (S). In mutant mice, most of the entorhinal cortex and, presumably, pre- and parasubiculum have merged into a small poorly defined region (?). The dentate gyrus and its dendritic layer (arrowhead) appear to be shrunken. The synaptic fields of the suprapyramidal mossy fibre layer (1) is reduced (arrow). In contrast, the intra-infrapyramidal mossy fibre projection (2) is enlarged (star), and has ectopic terminations near the fimbria. **g, h**, Coronal sections through the neocortex showing a dramatic reduction in the number of calretinin-positive cells. Scale bar: **a, c**, 500 μ m; **b, d**, 420 μ m; **e, f**, 0.5 mm; **g, h**, 230 μ m.

Although *tlx* transcripts are restricted to the ventricular zone during development, the size and number of cells in this region was not significantly different between wild-type and mutant animals up to embryonic day (E)14. At E15.5, however, mutant animals show a reduction in the number of cells in the intermediate zone of the cortex (Fig. 5a, b), consistent with a role for the *tlx* gene in ventricular zone proliferation and differentiation. In wild-type hippocampi, development of Ammons Horn is complete before birth, but only 15% of cells in the dentate gyrus are present⁶. In *tlx* mutant animals, the reduction in the number of granule cells in the dentate gyrus evident in adult brains was already apparent on day 4 (Fig. 5c–f). A reduction in the size of the olfactory bulb and rhinencephalic ventricular zone was also evident (data not shown). This suggests that the *tlx* protein may be required for the maintenance of the ventricular zone during later developmental stages. As some of these regions correspond to secondary proliferation zones of neuron precursors having left the germinal matrix⁶, the *tailless* protein may also play a role in controlling late migration and secondary proliferation.

Mutant animals were difficult to handle from birth, exhibiting hyperexcitability and very aggressive behaviour. All *tlx* male mutants mutilated or killed their littermates, although they were 30–50% smaller. Forty-five per cent of the females examined were also aggressive, a rare trait in female mice. Mutant females mated and gave birth normally but failed to care for their offspring. Initial attempts to obtain homozygous matings between mutant animals usually resulted in severe head/face and rump injuries and death in 80% of the females. When paired with non-oestrous wild-type females, mutant male animals attacked females on average eight times more frequently than wild-type males (not shown).

Disruption of hippocampal circuitry and entorhinal cortex by lesions is associated with spatial learning and memory deficits^{7–9}. Predictably, mutant animals also showed residual spatial learning abilities, as judged by the Morris water-maze test (not shown).

Emotions are thought to be governed by distinct brain regions¹⁰. Lesion studies in humans and other animals have indicated that the limbic system is important in fear, anger and sexual aggression^{10–12}. Such extreme aggressiveness and hyperexcitability is reminiscent of the 'septal rage' syndrome observed after lesions of the septum¹³.

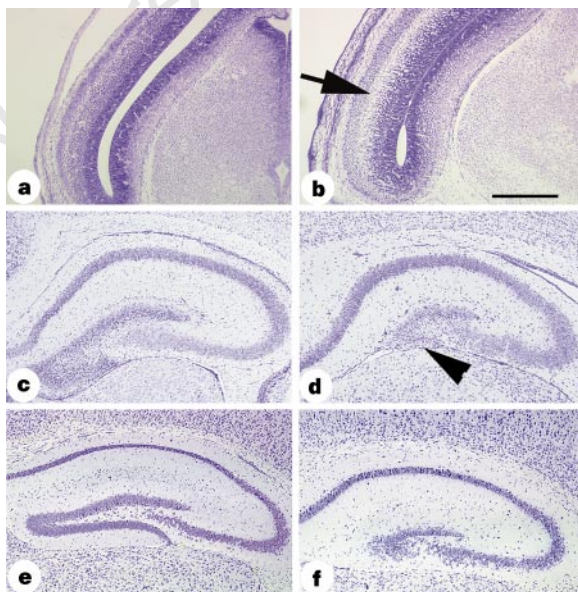


Figure 5 Developmental abnormalities in *tlx* mutant mice. Sections of wild-type (a, c, e) and mutant (b, d, f) brains. a, b, Coronal E15.5 brains show a reduction in cell number in the intermediate zone (arrow). c, d, In the dentate gyrus (arrowhead) of 4-day-old mice, cell numbers are reduced. e, f, Adult hippocampus, showing smaller dentate gyrus. Scale bar: a, c, 200 μ m; b, d, 166 μ m; e, f, 230 μ m.

The *tlx* phenotype cannot be due to olfactory deafferentation, which reduces or abolishes intermale aggression in mice¹⁴, but rather seems to reflect a massive disinhibition syndrome¹⁵ caused by missing or misrouted subcortical or limbic afferents to the hypothalamus¹⁶ or olfactory structures¹⁴. In summary, *tlx* is required for the maintenance of the ventricular zone after E15 and disruption of the *tlx* locus leads to impaired development of a specific subset of forebrain-derived structures. It is likely that the *tlx* gene is required for the proliferation or survival of a subpopulation of neural progenitors in the frontal cortex, amygdala, entorhinal cortex and dentate gyrus. Further analysis of the *tlx* mutation will clarify how this transcription factor can promote generation and survival of specific brain structures. □

Methods

Generation of *tlx*-deficient mice. The mouse *tlx* gene was cloned and mapped from an E14 genomic library¹⁷. The targeting vector contained 2.2 kb of 5' homology and 6 kb of 3' homology. The β -galactosidase PGK-neomycin cassette¹⁸ was fused to the first 14 amino acids of the *tailless* protein. We generated four chimaeric male mice from two independent 129/J ES cell clones, which transmitted the mutated allele when crossed to C57B1/6. F₁ animals (+/–) were mated *inter se* to produce homozygous mutant F₂ animals. For histological and behavioural analysis, F₂ animals from both clones were analysed separately.

Immunohistochemistry. Antibody staining was done as described¹⁹. Antibody sources: calretinin was from Sivant, Switzerland; Krox-24 was a gift from R. Bravo.

Received 24 February; accepted 3 September 1997.

- Pignoni, F. *et al.* The *Drosophila* gene *tailless* is expressed at the embryonic termini and is a member of the steroid receptor superfamily. *Cell* **62**, 151–163 (1990).
- Weigel, D., Jurgens, G., Kilger, M. & Jäckle, H. Two gap genes mediate maternal terminal pattern information in *Drosophila*. *Science* **248**, 495–498 (1990).
- Yu, R. T., McKeown, M., Evans, R. M. & Umeson, K. Relationship between *Drosophila* gap gene *tailless* and a vertebrate nuclear receptor Tlx. *Nature* **370**, 375–378 (1994).
- Monaghan, A. P., Grau, E., Bock, D. & Schuetz, G. The mouse homolog of the orphan nuclear receptor *tailless* is expressed in the developing forebrain. *Development* **121**, 839–853 (1995).
- Mangelsdorf, D. J. *et al.* The nuclear receptor superfamily: The second decade. *Cell* **83**, 835–839 (1995).
- Reznikov, K. Y. Cell proliferation and cytogenesis in the mouse hippocampus. *Adv. Anat. Embryol. Cell Biol.* **122**, 1–74 (1991).
- Keefe, J. O. & Nadel, L. *The Hippocampus as a Cognitive Map* (Clarendon, London, 1978).
- Morris, R. G. M. *et al.* Place navigation impaired in rats with hippocampal lesions. *Nature* **297**, 681–683 (1982).
- Morris, R. G. M. Development of the watermaze procedure for studying spatial learning in the rat. *J. Neurosci. Meth.* **11**, 47–60 (1984).
- Kandel, E. R., Schwartz, J. H. & Jessell, T. *Essentials of Neural Science and Behavior* (Prentice Hall International, London, 1995).
- Davis, M. The role of the amygdala in fear and anxiety. *Annu. Rev. Neurosci.* **15**, 353–357 (1994).
- Gray, J. A. *The Neuropsychology of Anxiety: an Enquiry into the Functions of the Septo-Hippocampal System* (Clarendon, London, 1982).
- Goodlet, C. R., Engellner, W. J., Burright, R. G. & Donovick, P. J. Influence of environmental rearing history and post surgical environmental change on the septal rage syndrome in mice. *Physiol. Behav.* **28**, 1077–1081 (1982).
- Edwards, D. A., Nahai, F. R. & Wright, P. Pathways linking the olfactory bulbs with the medial peptidic anterior hypothalamus are important for intermale aggression in mice. *Physiol. Behav.* **53**, 611–615 (1993).
- Albert, D. J. & Walsh, M. L. Neural systems and the inhibitory modulation of agonistic behaviour: a comparison of mammalian species. *Neurosci. Biobehav. Rev.* **8**, 5–24 (1984).
- Gilad, G. M. & Reis, D. J. Collateral sprouting in central mesolimbic dopamine neurons: biochemical and immunocytochemical evidence of changes in the activity and distribution of tyrosine hydroxylase in terminal fields and in cell bodies of A10 neurons. *Brain Res.* **160**, 17–26 (1979).
- Kaestner, K. H., Hiemisch, H., Luckow, B. & Schütz, G. The HNF-3 gene family of transcription factors in mice: gene structure, cDNA sequence, and mRNA distribution. *Genomics* **20**, 377–385 (1994).
- Kaestner, K. H. *et al.* Universal beta-galactosidase cloning vectors for promoter analysis and gene targeting. *Gene* **148**, 67–70 (1994).
- Gass, P., Herdegen, T., Bravo, R. & Kiessling, M. Induction of immediate early gene encoded proteins in the rat hippocampus after bicuculline-induced seizures: differential expression of Krox-24, FOS and Jun proteins. *Neuroscience* **48**, 315–324 (1992).
- Tamura, T. *et al.* Striking homology of the 'variable' N-terminal as well as the 'conserved core' domains of the mouse and human TATA-factors (TFIID). *Nucleic Acids Res.* **19**, 3861–3865 (1991).

Acknowledgements. We thank E. Grau, E. Schmid, S. Ridder, H. Glaser and A. Klewe-Nebenius for technical assistance; J. Blendy, H. Kaestner and M. Nichols for critically reading the manuscript; R. Lang and M. Stagliar-Bozicevic for help with behavioural analyses; V. Fleischer for help with photography; and M. Bock for secretarial assistance. This work was supported by the Deutsche Forschungsgemeinschaft, the European Community, BMBF (Forschungszentrum Jülich/Schering), Fonds der Chemischen Industrie, by the Swiss National Foundation for Scientific Research and the Human Science Frontier Programme (to H.P.L.).

Correspondence and requests for materials should be addressed to G.S. (e-mail: schuetz@dkfz-Heidelberg.de).

Activation of peripheral CB₁ cannabinoid receptors in haemorrhagic shock

Jens A. Wagner*, Károly Varga*, Earl F. Ellis, Beverly A. Rzigalinski, Billy R. Martin & George Kunos

Department of Pharmacology and Toxicology, Medical College of Virginia, Virginia Commonwealth University, PO Box 980613, Richmond, Virginia 23298, USA

* These authors contributed equally to this work

Anandamide, an endogenous cannabinoid ligand¹, binds to CB₁ cannabinoid receptors² in the brain and mimics the neurobehavioural actions of marijuana^{3,4}. Cannabinoids and anandamide also elicit hypotension mediated by peripheral CB₁ receptors^{5–8}. Here we report that a selective CB₁ receptor antagonist, SR141716A⁹, elicits an increase in blood pressure in rats subjected to haemorrhagic shock, whereas similar treatment of normotensive rats or intracerebroventricular administration of the antagonist during shock do not affect blood pressure. Blood from haemorrhaged rats causes hypotension in normal rats, which can be prevented by SR141716A but not by inhibition of nitric oxide synthase in the recipient. Macrophages and platelets from haemorrhaged rats elicit CB₁ receptor-mediated hypotension in normotensive recipients, and incorporate arachidonic acid or ethanolamine into a product that co-elutes with anandamide on reverse-phase high-performance liquid chromatography. Also, macrophages from control rats stimulated with ionomycin or bacterial phospholipase D produce anandamide, as identified by gas chromatography and mass spectrometry. These findings indicate that activation of peripheral CB₁ cannabinoid receptors contributes to haemorrhagic hypotension, and anandamide produced by macrophages may be a mediator of this effect.

Rats anaesthetized with urethane were subjected to standardized, stepwise bleeding until mean blood pressure stabilized at 40 mm Hg (refs 10, 11). Blood pressure remained stable at this level for 25–35 min, after which it started to decline and the animal died. At the beginning of the stable hypotensive phase, some rats received 3 mg kg^{–1} of SR141716A intravenously (i.v.), which caused a marked and prolonged pressor response that lasted 20–30 min (Fig. 1). The

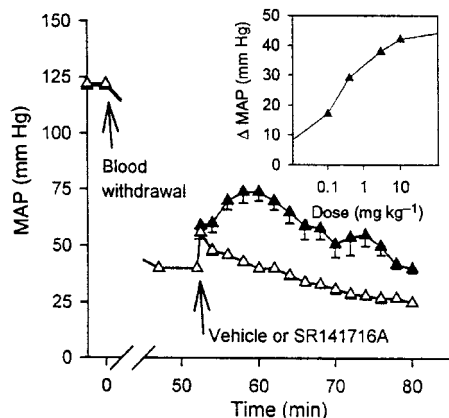


Figure 1 The effects of SR141716A (3 mg kg^{–1} i.v. in 0.2 ml, ▲, *n* = 7) or vehicle (0.2 ml, △, *n* = 7) on blood pressure in rats subjected to haemorrhagic shock. Inset: dose-dependence of the pressor effect of SR141716A in shock, each point representing the mean of 3–7 experiments. The ED₅₀ value for the pressor effect of SR141716A was calculated by using the ALLFIT program. Vertical bars in Figs 1–3 represent s.e.m.

dose of SR141716A that caused a half-maximal pressor effect (0.37 ± 0.08 mg kg^{–1}, Fig. 1, inset) was similar to the dose that caused half-maximal inhibition of anandamide-induced hypotension (0.29 ± 0.14 mg kg^{–1}) (ref. 8) or Δ^9 -tetrahydrocannabinol (THC)-induced neurobehavioural effects in rats (0.11 – 0.54 mg kg^{–1}) (ref. 9), which supports the involvement of CB₁ receptors in shock-related hypotension. Other rats received vehicle, which only caused a small and transient pressor effect (Fig. 1). In control animals, the same dose of SR141716A caused no significant change in blood pressure either when injected during normotensive conditions (-6 ± 3 mm Hg, *n* = 6), or after mean blood pressure had been reduced to 53 ± 3 mm Hg due to α -adrenergic blockade by phentolamine, 2 mg kg^{–1} i.v. (*n* = 3, not shown). Furthermore, in three rats subjected to haemorrhagic hypotension, injection of 300 μ g kg^{–1} SR141716A into the cisterna magna or the IVth cerebral ventricle failed to influence blood pressure (not shown).

The specificity of the pressor effect of SR141716A during haemorrhagic hypotension is further indicated by the ability of a CB₁ receptor agonist to reverse it. WIN-55212-2 is a potent CB₁ receptor agonist¹², and it causes pronounced hypotension in urethane-anaesthetized control rats with a half-maximal effect at 0.1 mg kg^{–1} i.v. (ref. 8). In three rats in haemorrhagic shock, cumulative doses of WIN-55212-2 were given i.v. at the plateau of the pressor response elicited by 3 mg kg^{–1} SR141716A. WIN-55212-2 started to reverse the pressor response to SR141716A at a dose of 1 mg kg^{–1}, and rapid reversal to 40 mm Hg was achieved with a cumulative dose of 20 mg kg^{–1}. This dose range is similar to that needed to lower blood pressure in normal rats pretreated with the same dose of SR141716A (not shown), which suggests that the pressor effect of SR141716A in shock and its reversal by WIN-55212-2 are mediated by the same receptor. Taken together, these findings suggest that activation of peripheral CB₁ receptors by an endogenous ligand contributes to haemorrhagic hypotension.

In order to test this possibility further, we examined the effects of exchange transfusion of blood from heparinized donor rats on the blood pressure of recipient rats. Transfusion of 3 ml of blood from normal donor to normal recipient rats had no effect on blood pressure. However, in five rats, transfusion of the same volume of blood obtained following haemorrhage ('shock' blood) gradually lowered blood pressure to a plateau 33 ± 9 mm Hg below preinfusion levels, and the hypotension lasted more than 2 h. This effect was completely blocked by pretreatment of recipients with 3 mg kg^{–1} SR141716A (peak effect of 'shock' blood: $+5 \pm 7$ mm Hg, *n* = 3, *P* < 0.005).

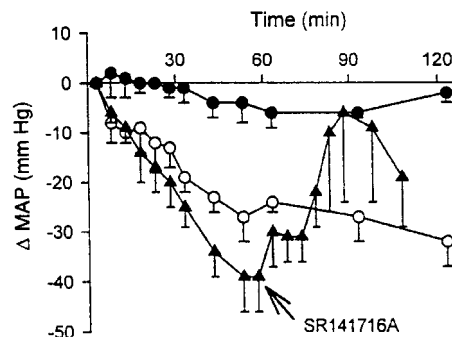


Figure 2 SR141716A inhibits the hypotensive response of control rats to i.v. administered mononuclear cells isolated from the blood of haemorrhaged rats. Recipient rats were either untreated (○, *n* = 4), pretreated with SR141716A, 3 mg kg^{–1} i.v. 10 min before receiving the injection of mononuclear cells (●, *n* = 3), or treated with the same dose of SR141716A after injection of mononuclear cells, as indicated by the arrow (▲, *n* = 4).

Plasma obtained by centrifugation of 3 ml of 'shock' blood caused only a minimal change in blood pressure (-6 ± 4 mm Hg, $n = 3$) when injected into recipient rats. In contrast, exchange transfusion of the cellular pellet resuspended to the original blood volume in saline caused marked and prolonged hypotension (>2 h, -30 ± 8 mm Hg, $n = 5$), which could be completely blocked by pretreatment of the recipients with 3 mg kg^{-1} SR141716A ($+8 \pm 2$ mm Hg, $n = 3$, $P < 0.005$). Injection of blood cells from

normotensive donors caused minimal change in blood pressure (-8 ± 4 mm Hg, $n = 3$). Cells from 3 ml of blood were separated into mononuclear, granulocyte, and erythrocyte fractions, and each cell type resuspended in 1 ml of normal saline. Intravenous injection of the erythrocyte fraction did not affect blood pressure, whereas the granulocytes caused a transient (<1 min) hypotension and tachycardia, which was similar whether cells were obtained from normotensive or haemorrhaged donors, and was unaffected by SR141716A pretreatment. Injection of mononuclear cells from normotensive donors caused only minimal hypotension (-7 ± 1 mm Hg, $n = 7$), whereas the same cells from haemorrhaged rats caused prolonged hypotension that could be prevented by pretreatment of the recipients with 3 mg kg^{-1} of SR141716A (Fig. 2). Also, SR141716A was able to reverse the hypotension when administered following administration of the mononuclear fraction (Fig. 2).

Several lines of evidence suggest that changes in NO synthase activity and/or in circulating levels of nitrosothiols¹³ cannot explain the CB₁ receptor-mediated hypotension observed in shock, although some contribution of the NO system cannot be definitively excluded. First, inhibition of endothelial NO synthase in recipient rats by pretreatment with 20 mg kg^{-1} nitro-L-arginine methyl ester (L-NAME)¹⁴ caused a rapid and sustained (>2 h) increase in basal blood pressure by 20 ± 4 mm Hg, but did not prevent the hypotensive response to 'shock' blood (peak change: -34 ± 12 mm Hg, $n = 4$). Second, we found no significant difference between control and 'shock' cells in the rate of nitrite (1.32 ± 0.16 versus $1.36 \pm 0.15 \text{ nmol h}^{-1}$ per 10^6 macrophages) and nitrate production (3.56 ± 0.61 versus $4.47 \pm 1.04 \text{ nmol h}^{-1}$ per 10^6 macrophages, $n = 4$) during the first 3 h following their isolation, which agrees with a previous report¹⁵. Third, pretreatment of donor animals with 20 mg kg^{-1} L-NAME before haemorrhage did not block the transfer of hypotensive activity by macrophages (not

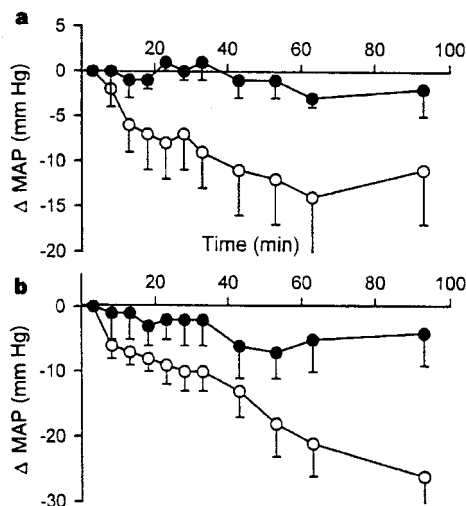


Figure 3 SR141716A inhibits the hypotensive response of control rats to i.v. injection of mononuclear cells from platelet-depleted blood (a) or of platelet-rich plasma (b) obtained from animals in shock. Symbols as in Fig. 1, $n = 3-5$ (a) or 7-12 (b).

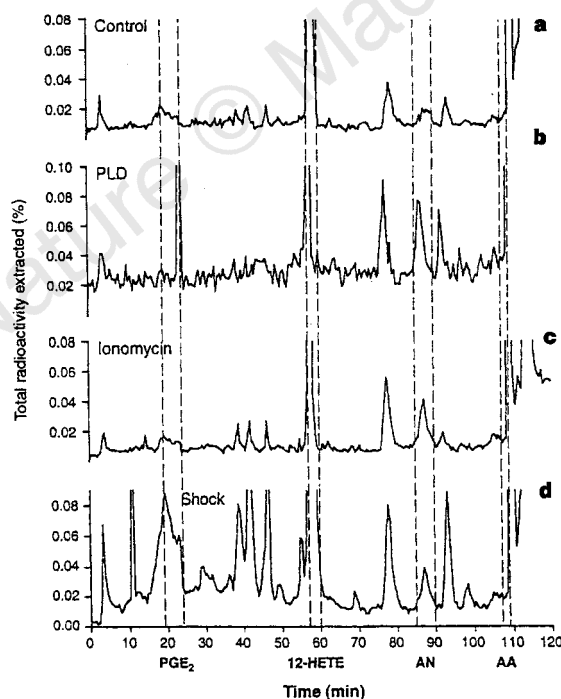


Figure 4 HPLC analysis of ^3H -arachidonic acid metabolites generated *in vitro* by macrophages isolated from rat blood. Cells isolated from a normotensive (a to c) and a haemorrhaged rat (d) were plated, preincubated with ^3H -arachidonic acid for 16 h (a to c) or 2 h (d), and then treated with vehicle (a and d), phospholipase D (b) or ionomycin (c) as described under Methods. Elution times of radiolabelled standards are indicated by double dashed lines. The radioactivity (d.p.m.) in the peaks coeluting with anandamide was 370 (a), 1,491 (b) and 899 (c). In another similar preparation the corresponding values were 295 (a), 785 (b) and 1,029 (c).

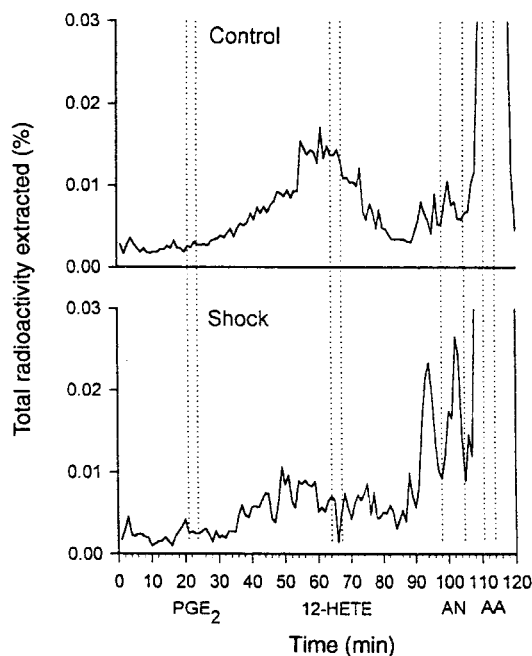


Figure 5 HPLC analysis of ^3H -ethanolamine metabolites generated *in vitro* by macrophages isolated from 3 ml blood from a control (top) or a haemorrhaged rat (bottom). Cells were preincubated for 16 h with ^3H -ethanolamine, followed by 30 min in serum-free medium containing PMSF and unlabelled anandamide (see Methods). Processing and analysis of the samples was as in Fig. 4, except for using a different HPLC machine and column. This is one of three experiments with similar results.

shown). Fourth, the hypotensive response to the i.v. injection of $5 \mu\text{g kg}^{-1}$ sodium nitroprusside, a nitric oxide donor, was the same before and after the administration of 3 mg kg^{-1} SR141716A (-57 ± 5 versus -63 ± 5 mm Hg, $n = 3$).

The mononuclear fraction was further separated by allowing the macrophages and most of the contaminating platelets ($10\text{--}20$ platelets per macrophage) to adhere to a plastic culture dish. Intravenous injection of the non-adherent cells, mainly lymphocytes, did not significantly alter the blood pressure of recipient rats. To test the effect of macrophages and platelets separately, platelet-rich plasma was prepared from 3 ml of whole blood, yielding $8\text{--}10 \times 10^8$ pure platelets. The platelets were removed and replaced with saline to prepare the mononuclear fraction as before, which now contained $5\text{--}6 \times 10^6$ macrophages plus $2\text{--}5$ platelets per macrophage. Injection of either the platelet-rich plasma (Fig. 3a) or the platelet-depleted monocyte fraction (Fig. 3b) from haemorrhaged rats into normal recipients caused prolonged hypotension that could be prevented by pretreatment of the recipient with SR141716A. Injection of either fraction obtained from normal donors had no effect on blood pressure (not shown). These findings indicate that both macrophages and platelets contribute to the hypotensive activity of the mononuclear cell fraction.

Endorphins are known to contribute to hypotension in various forms of shock through an action at central opiate receptors, and naloxone was reported to inhibit the hypotension as well as to improve survival¹¹. Unexpectedly, pretreatment of rats

with 3 mg kg^{-1} SR141716 before bleeding reduced the survival time (measured from the last blood removal before blood pressure stabilized at 40 mm Hg for at least 5 min) from 30 ± 5 min ($n = 9$) to 8 ± 1 min ($n = 5$, $P < 0.005$). In contrast, pretreatment with 1 mg kg^{-1} of THC, or $2 \mu\text{g kg}^{-1}$ of HU-210, a potent hypotensive cannabinoid devoid of an initial pressor effect⁸, significantly increased survival to 51 ± 8 ($n = 5$, $P < 0.05$) or 58 ± 8 min ($n = 5$, $P < 0.05$), respectively. This suggests that, in contrast to the effect of endogenous opioids, activation of CB₁ receptors is beneficial for survival in shock. Although the underlying mechanism is unclear, a favourable redistribution of cardiac output or improved microcirculation by localized vasodilation^{16,17} are possibilities.

Anandamide has been shown to be formed in and released from neurons in a calcium-dependent manner by phospholipase D-mediated cleavage of *N*-arachidonoyl phosphatidylethanolamine¹⁸. A similar process has been documented in a macrophage cell line^{19,20}. We tested whether anandamide was formed in macrophages obtained from circulating blood. The cells were maintained in primary culture and preincubated with ^3H -arachidonic acid. Reverse phase high-performance liquid chromatography (HPLC) separation of radioactive products generated by unstimulated cells from control rats yielded a small, composite peak coeluting with authentic anandamide (Fig. 4a). However, a much larger and more distinct peak emerged in the same position in aliquots of the same cells briefly exposed to bacterial phospholipase D (Fig. 4b) or to the calcium ionophore, ionomycin (Fig. 4c), agents that have been shown to release anandamide from a macrophage cell line^{19,20}. A similar peak also increases in size in cells isolated from a rat in haemorrhagic shock (Fig. 4d). Because other arachidonate products may coelute with anandamide in this HPLC system²¹, in additional experiments we incubated the macrophage/platelet preparation with ^3H -ethanolamine and then separated the products by reverse phase HPLC. A radiolabelled product coeluting with authentic anandamide appeared in cells from haemorrhaged rats, and the size of this peak was much smaller in the same number of cells from control rats (Fig. 5). This double labelling strongly suggests that the product generated is anandamide. We next examined the ability of the same cells to generate anandamide in the absence of exogenous precursors. A macrophage/platelet preparation stimulated *in vitro* by ionomycin plus PLD generated a product that could be identified as anandamide by gas chromatography/mass spectrometry, based on its retention time and *m/z* ratios (compare Fig. 6a and b). When a 10 times larger number of pure platelets was used in a similar experiment, the peak at the retention time of anandamide was much smaller and could not be clearly identified as anandamide, based on *m/z* ratios (not shown). This suggests that the major source of anandamide in Fig. 6 was macrophages.

Most of the anandamide formed and released by neurons or a macrophage cell line remains cell-associated because of a high-affinity uptake system^{18,19}. A similar behaviour of circulating macrophages may explain the lack of hypotensive activity found in the plasma of animals in shock. Shock is known to be associated with sludging of blood cells and increased direct contact between macrophages, platelets and the vascular wall¹⁰. Endothelial or vascular smooth muscle cells may therefore be the site of the hypotensive action of macrophage-derived anandamide. However, platelets from rats in shock also elicited CB₁ receptor-mediated hypotension, yet produced much less, if any, anandamide. Autologous platelets have been reported to greatly enhance cytokine production by mononuclear cells²², and platelets or platelet-derived factors may similarly induce anandamide production in activated macrophages. The lack of 'tonic' activity of this CB₁ receptor-mediated mechanism and its persistence after blockade of endothelial NO synthase distinguishes it from NO-mediated vasodilation. How this new vasodilator mechanism is activated in haemorrhagic shock, remains to be determined.

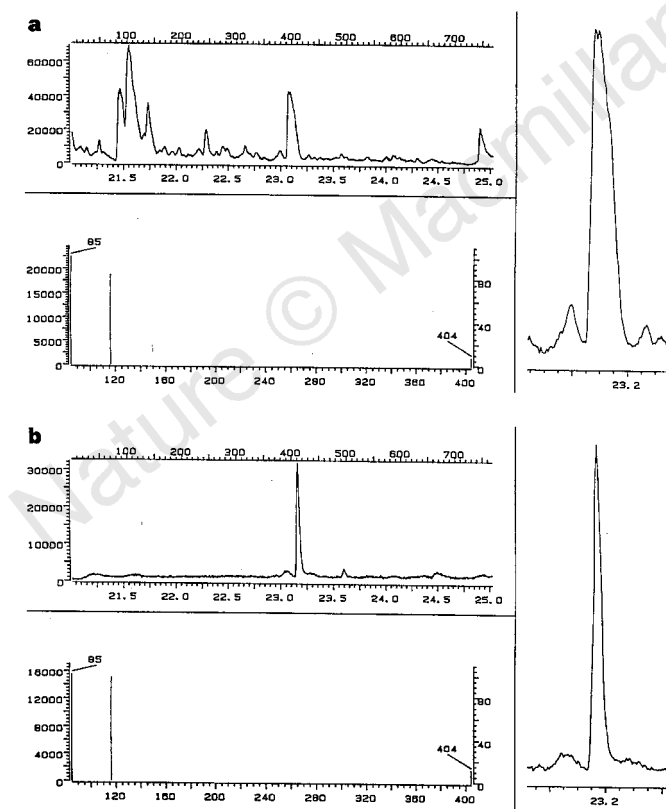


Figure 6 Identification by GC/MS of anandamide generated by macrophages and platelets. **a**, 5×10^7 macrophages plus $5\text{--}10 \times 10^8$ platelets isolated from 260 ml control rat blood were incubated for 20 min in serum-free medium containing $200 \mu\text{M}$ PMSF, 10 U ml^{-1} PLD and $5 \mu\text{M}$ ionomycin. Cells plus medium were extracted with ethylacetate, dried and resuspended in ethanol for separation by reverse phase HPLC. Fractions corresponding to the elution profile of anandamide were pooled and subjected to GC/MS (see Methods). **b**, GC/MS analysis of 100 ng anandamide. Upper left panels, combined currents obtained by selected ion monitoring on *m/z* 85, *m/z* 116 and *m/z* 404; lower left panels: ratio of the monitored ions at the apex of the anandamide peak; right panels: expanded view of the anandamide region of the ion chromatogram.

Methods

Haemorrhagic shock. Male Sprague–Dawley rats (300–350 g) were anaesthetized with urethane, 0.7 g kg^{-1} i.v. + 0.3 g kg^{-1} i.p., heparinized (500 IU kg^{-1}), and kept on a 37°C heating pad. Polyethylene cannulae were placed into both femoral arteries for monitoring arterial blood pressure and for withdrawal of blood, respectively, and into the femoral vein for drug injections. Rats were then subjected to gradual bleeding, in steps of 0.5 ml gradually tapered to 0.1 ml , until mean arterial pressure (MAP) stabilized at 40 mm Hg within $40\text{--}45 \text{ min}$ of withdrawal of a total of $3.1 \pm 0.1 \text{ ml}$ blood per 100 g body weight (52% of the total blood volume²³).

Isolation of blood cells. Cells and plasma were separated by centrifugation. The combined cells were washed three times and resuspended to the original volume in saline. Erythrocyte, granulocyte and mononuclear cell fractions were prepared from whole blood by ficoll/isopaque density gradient centrifugation, as described by Böyum²⁴. Erythrocytes were separated from the granulocytes by sedimentation through Dextran 500 (ref. 24). Platelet-rich plasma was prepared by centrifugation of whole blood at 64 g for 15 min .

Measurement of nitrite and nitrate production. Macrophages isolated from control and haemorrhaged rats (10^6 cells plus contaminating platelets) were plated and maintained in 1 ml Krebs buffer containing 10 mM glucose. The buffer was removed at 1, 2 and 3 h and replaced with fresh buffer. Nitrite concentrations were determined by the Griess reaction²⁵ in duplicate aliquots of the buffer, in one of which nitrates were first reduced using the cadmium method²⁵. Nitrate concentrations were calculated as the difference between the two values.

HPLC analyses. Mononuclear cells were isolated and plated on plastic Petri dishes in 4 ml RPMI 1640 medium containing 10% fetal calf serum, $10 \mu\text{g ml}^{-1}$ gentamycin and $10 \mu\text{g ml}^{-1}$ fungizone. After 60 min , lymphocytes were removed by aspiration, and the adherent macrophages ($5\text{--}6 \times 10^6$ cells per dish) plus contaminating platelets ($\sim 10^8$ cells per dish) were preincubated with ^3H -arachidonic acid (220 Ci mmol^{-1} , $5 \mu\text{Ci ml}^{-1}$) or ^3H -ethanolamine (18 Ci mmol^{-1} , $10 \mu\text{Ci ml}^{-1}$), for 2 or 16 h , as indicated. Following preincubation, the cells were washed 4 times with serum-free RPMI 1640 medium containing $200 \mu\text{M}$ phenylmethylsulphonylfluoride (PMSF) and $100 \mu\text{M}$ unlabelled anandamide to prevent the breakdown and cellular reuptake of anandamide formed by the cells, respectively¹⁸. The cells were then treated for 20 min with vehicle or phospholipase D (PLD) from *Streptomyces chromofuscus* (10 U ml^{-1}), or for 5 min with ionomycin ($5 \mu\text{M}$), as indicated. At the end of the treatment the cells plus medium were extracted 3 times with 2 vol ethyl acetate, the organic phase was dried under nitrogen, resuspended in $400 \mu\text{l}$ methanol/water (3:1) and analysed by reverse phase HPLC on a Radial-Pak C₁₈ column. Samples were eluted at a flow rate of 1 ml min^{-1} , using a continuous gradient of water/acetic acid (100:0.05, vol/vol) and methanol/acetic acid (100:0.05, vol/vol), as described²¹. Fractions (0.5 ml) were collected and radioactivity determined by liquid scintillation spectrometry.

Gas chromatography/mass spectrometry. Ethylacetate extracts of macrophages plus contaminating platelets were prepurified by reverse phase HPLC. The fractions corresponding to the elution profile of anandamide were pooled, evaporated to dryness under nitrogen, and derivatized with bis(trimethylsilyl)trifluoroacetamide (BSTFA) for 1 h at 70°C . Samples were then dried again and resuspended in *n*-hexane. The trimethylsilyl ether (TMS) derivatives were injected in 'splitless' mode into a Hewlett-Packard 5890 gas chromatograph (GC) equipped with an HP-1 capillary column ($25 \text{ m} \times 0.2 \text{ mm} \times 0.33 \mu\text{m}$) and interfaced with a Hewlett-Packard 5988A mass spectrometer. Four minutes after injection the oven temperature was ramped from 90°C to 300°C at $10^\circ\text{C min}^{-1}$. Helium was used as the carrier gas. The injector and detector temperatures were maintained at 260°C , and the source temperature was 200°C . The TMS derivative of anandamide eluted from the GC at $\sim 23.1 \text{ min}$. Based on the electron impact spectrum, selected ion monitoring analyses were performed on m/z 85, m/z 116 and m/z 404 (M-15).

Drugs. The vehicle for SR141716A [*N*-(piperidyl-1-yl)-5-(4-chlorophenyl)-(2,4-dichloro-phenyl)-4-methyl-1H-pyrazol-3-carboxamide HCl]⁹ or anandamide (arachidonyl ethanolamide) was emulphor:ethanol:saline 1:1:18. Emulphor is a polyoxyethylated vegetable oil.

Received 3 July; accepted 12 September 1997.

- Devane, W. A. *et al.* Isolation and structure of a brain constituent that binds to the cannabinoid receptor. *Science* **258**, 1946–1948 (1992).

- Matsuda, L. A., Lolait, S. J., Brownstein, M. J., Young, A. C. & Bonner, T. I. Structure of a cannabinoid receptor and functional expression of the cloned cDNA. *Nature* **346**, 561–564 (1990).
- Fride, E. & Mechoulam, R. Pharmacological activity of the cannabinoid receptor agonist, anandamide, a brain constituent. *Eur. J. Pharmacol.* **231**, 313–314 (1993).
- Smith, P. B. *et al.* The pharmacological activity of anandamide, a putative endogenous cannabinoid, in mice. *J. Pharmacol. Exp. Ther.* **270**, 219–227 (1994).
- Varga, K., Lake, K. D., Martin, B. R. & Kunos, G. Novel antagonist implicates the CB1 cannabinoid receptor in the hypotensive action of anandamide. *Eur. J. Pharmacol.* **278**, 279–283 (1995).
- Varga, K., Lake, K. D., Huangfu, D., Guyenet, P. G. & Kunos, G. Mechanism of the hypotensive action of anandamide in anesthetized rats. *Hypertension* **28**, 682–688 (1996).
- Lake, K. D., Martin, B. R., Kunos, G. & Varga, K. Cardiovascular effects of anandamide in anesthetized and conscious normotensive and hypertensive rats. *Hypertension* **29**, 1204–1210 (1997).
- Lake, K. D., Varga, K., Compton, D. R. & Kunos, G. Cannabinoid-induced hypotension and bradycardia in rats is mediated by CB1 cannabinoid receptors. *J. Pharmacol. Exp. Ther.* **281**, 1030–1037 (1997).
- Rinaldi-Carmona, M. *et al.* SR141716A, a potent and selective antagonist of the brain cannabinoid receptor. *FEBS Lett.* **350**, 240–244 (1994).
- Peitzmann, A. B. *et al.* Hemorrhagic shock. *Curr. Probl. Surg.* **32**, 927–1002 (1995).
- Holaday, J. W. Cardiovascular effects of endogenous opiate systems. *A. Rev. Pharmacol. Toxicol.* **23**, 541–594 (1983).
- Felder, C. C., Veluz, J. S., Williams, H. L., Briley, E. M. & Matsuda, L. A. Cannabinoid agonists stimulate both receptor- and non-receptor-mediated signal transduction pathways in cells transfected with an expressing cannabinoid receptor clones. *Mol. Pharmacol.* **42**, 838–845 (1992).
- Myers, P. R., Minor, R. L., Guerra, R., Bates, J. N. & Harrison, D. G. Vasorelaxant properties of the endothelium-derived relaxing factor more closely resemble N-nitrosocysteine than nitric oxide. *Nature* **345**, 161–163 (1990).
- Knowles, R. G. & Moncada, S. Nitric oxide synthases in mammals. *Biochem. J.* **298**, 249–258 (1994).
- Naziri, W. *et al.* Hemorrhagic shock-induced alterations in circulating and bronchoalveolar macrophage nitric oxide production. *J. Surg. Res.* **59**, 146–152 (1995).
- Ellis, E. F., Moore, S. S. & Willoughby, K. A. Anandamide and Δ^9 -THC dilation of cerebral arterioles is blocked by indomethacin. *Am. J. Physiol.* **269**, H1859–H1864 (1995).
- Gebremedhin, G., Campbell, W. B., Hillard, C. J. & Harder, D. R. Inhibitory effects of anandamide on Ca^{2+} currents in cat cerebral arterial smooth muscle cells. *FASEB J.* **10**, A302 (1996).
- Di Marzo, V. *et al.* Formation and inactivation of endogenous cannabinoid anandamide in central neurons. *Nature* **372**, 686–691 (1994).
- Di Marzo, V., De Petrocellis, L., Sepe, N. & Buono, A. Biosynthesis of anandamide and related acylethanolamides in mouse J774 macrophages and N18 neuroblastoma cells. *Biochem. J.* **316**, 977–984 (1996).
- Bisogno, T., Maurelli, S., Melck, D., De Petrocellis, L. & Di Marzo, V. Biosynthesis, uptake, and degradation of anandamide and palmitoylethanolamide in leukocytes. *J. Biol. Chem.* **272**, 3315–3323 (1997).
- Amrutesh, S. C., Falck, J. R. & Ellis, E. F. Brain synthesis and cerebrovascular action of epoxyeic acid metabolites of arachidonic acid. *J. Neurochem.* **58**, 503–510 (1992).
- Aiura, K. *et al.* Interaction with autologous platelets multiplies interleukin-1 and tumor necrosis factor production in mononuclear cells. *J. Inf. Dis.* **175**, 123–129 (1997).
- Collins, J. A., Braitberg, A., Margraf, H. W. & Butcher, H. R. Hemorrhagic shock in rats. *Arch. Surg.* **99**, 484–488 (1969).
- Böyum, A. Isolation of mononuclear cells and granulocytes from human blood. *Scand. J. Lab. Clin. Invest.* **21** (suppl. 97), 77–89 (1968).
- Cortas, N. K. & Wakid, N. W. Determination of inorganic nitrate in serum and urine by a kinetic cadmium-reduction method. *Clin. Chem.* **36**, 1440–1443 (1990).

Acknowledgements. We thank K. Willoughby for help with the HPLC assays, T. Bridgen for help with GC/MS, E. Ishac for preparing illustrations and J. Lowe for providing SR141716A. This work was funded by grants from the American Heart Association Virginia Affiliate and from the National Institutes of Health. J.A.W. was supported by a fellowship from the Deutsche Forschungsgemeinschaft.

Correspondence and requests for materials should be addressed to G.K. (e-mail: gkunos@hsc.vcu.edu).

Leptin inhibits hypothalamic neurons by activation of ATP-sensitive potassium channels

D. Spanswick, M. A. Smith, V. E. Groppi*, S. D. Logan & M. L. J. Ashford

Department of Biomedical Sciences, University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen AB25 2ZD, UK

* Pharmacia & Upjohn, 7000 Portage Road, Kalamazoo, Michigan 49001, USA

Leptin, the protein encoded by the obese (*ob*) gene, is secreted from adipose tissue and is thought to act in the central nervous system to regulate food intake and body weight^{1,2}. It has been proposed that leptin acts in the hypothalamus^{3–5}, the main control centre for satiety and energy expenditure⁶. Mutations in leptin or the receptor isoform (Ob-R_L) present in hypothalamic neurons result in profound obesity and symptoms of non-insulin-dependent diabetes^{7–10}. Here we show that leptin hyperpolarizes glucose-receptive hypothalamic neurons of lean Sprague–Dawley

and Zucker rats, but is ineffective on neurons of obese Zucker (*fa/fa*) rats. This hyperpolarization is due to the activation of a potassium current, and is not easily recovered on removal of leptin, but is reversed by applying the sulphonylurea, tolbutamide. Single-channel recordings demonstrate that leptin activates an ATP-sensitive potassium (K_{ATP}) channel. Our data indicate that the K_{ATP} channel may function as the molecular end-point of the pathway following leptin activation of the Ob- R_L receptor in hypothalamic neurons.

Leptin (4–50 nM) applied to hypothalamic slices decreased the firing rate of action potentials and input resistance of a subgroup of neurons, recorded from ventromedial ($n = 15$) and arcuate ($n = 3$) nuclei. These actions were accompanied by a slow and progressive hyperpolarization to a new equilibrium, 5–15 min after application (Fig. 1a). There was no obvious dose dependence for the action of leptin, as bath application induced an effect at all concentrations examined with the same slow time course with little or no reversibility over the duration of these experiments (up to 60 min). These effects were observed using whole-cell and intracellular recordings. The mean resting membrane potential before and 15 min after leptin application was -53 ± 2 and -73 ± 4 mV, respectively, with a mean peak hyperpolarization of 20.0 ± 3.0 mV ($n = 9$). The hyperpolarization was accompanied by a decrease in neuronal input resistance (Fig. 1b) by $41 \pm 4\%$ (range 36–48%, $n = 9$). Current-voltage relations before and after leptin application (Fig. 1c) showed that the conductance increase had a reversal potential of -88 ± 2 mV ($n = 3$), indicating an increase in potassium current (predicted reversal potential is -93 mV).

Hypothalamic neurons sensitive to leptin from both nuclei were also susceptible to changes in the extracellular concentration of glucose, in that a decrease resulted in hyperpolarization, decreased action-potential frequency and input resistance (see Fig. 2a). These neurons have been termed glucose-responsive (GR) neurons^{11,12}. GR neurons are also excited by the application of tolbutamide (Fig. 2a), particularly in the absence of extracellular glucose¹³. Application of 20 nM leptin causes hyperpolarization and an increased input conductance (Fig. 2b). Application of tolbutamide (50–200 μ M)

to GR neurons exposed to leptin (see Fig. 2b) reversed the effects of this hormone, inducing depolarization (Fig. 2c) and an increased input resistance to $69 \pm 16\%$ ($n = 3$) of the value observed following exposure to leptin. In the presence of tolbutamide and leptin, GR neurons often reached threshold for action-potential firing (Fig. 2b; $n = 6$). Removal of tolbutamide allowed the leptin response to re-emerge, even though the leptin was also washed out of the bath, as neurons re-established a hyperpolarized membrane potential with an associated increased conductance (data not shown). Control experiments on non-GR neurons (that is, neurons insensitive to a reduction of extracellular glucose concentration or application of tolbutamide) indicated that leptin (4–100 nM) had no effect on their resting membrane potential, conductance or firing frequency ($n = 65$), but induced a small 2–5 mV depolarization and increased firing frequency in four neurons.

The obese Zucker rat has a mutant Ob- R_L receptor isoform^{14,15}, which results in a much reduced signalling capability^{16,17}. In five GR neurons identified from the ventromedial nucleus of lean Zucker rats, leptin (10–20 nM) hyperpolarized the membrane potential from -48 ± 3 mV to -65 ± 6 mV with a concomitant decrease in input resistance of $41 \pm 13\%$ (Fig. 2c). The leptin-induced increase in conductance had a reversal potential of -90 ± 2 mV ($n = 3$). Tolbutamide (200 μ M) reversed the effects of leptin ($n = 3$), causing depolarization and increased input resistance (to $87 \pm 12\%$ of value in leptin). In other neurons examined, leptin had no effect ($n = 53$) or caused a small excitation ($n = 2$). Leptin (10–20 nM) had no effect on hypothalamic neurons ($n = 39$) recorded from obese Zucker rats, including three identified as GR neurons (Fig. 2d).

Reducing the extracellular glucose concentration bathing GR neurons or applying tolbutamide are considered to alter membrane potential and input resistance by activation and inhibition, respectively, of ATP-sensitive K^+ (K_{ATP}) channels^{12,13}. Therefore leptin activation was examined using single-channel recording techniques from hypothalamic slices ($n = 5$) and acutely dissociated neurons ($n = 38$), using non-identified cells in the presence of extracellular glucose, with no discernible difference in the results obtained. Bath application of leptin (10–100 nM) during cell-attached recordings

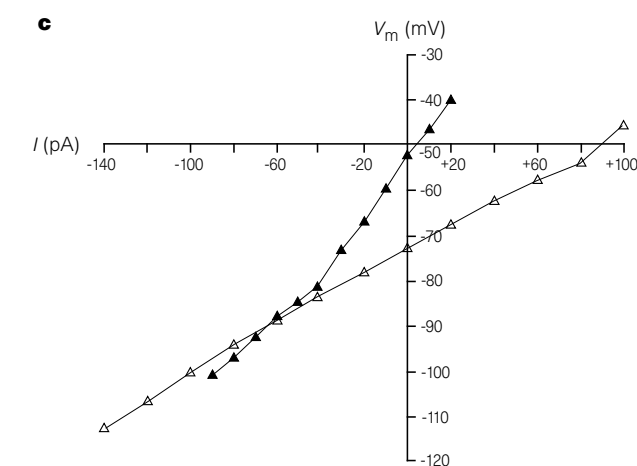
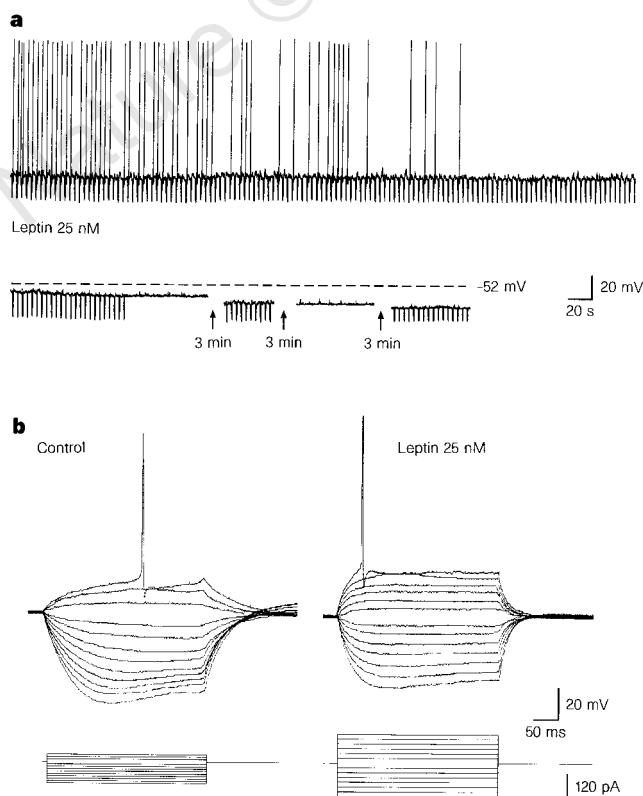


Figure 1 Leptin hyperpolarizes hypothalamic neurons by activation of a K^+ conductance. **a**, Continuous record illustrating leptin (25 nM) causing membrane hyperpolarization, from a resting membrane potential of -52 mV to -73 mV, and cessation of action-potential activity. Note the time breaks in the record, indicated by the arrows. Leptin decreased input resistance, indicated by the reduction in amplitude of the electrotonic potential. **b**, Superimposed electrotonic potentials (upper records) evoked in response to current pulse injection (below) in the absence and presence of 25 nM leptin. **c**, Corresponding current-voltage relation of data in **b**; control (filled triangles) and 25 nM leptin (open triangles).

from isolated neurons (no leptin in the electrode solution) did not lead to activation of potassium channels within the area of the patch ($n = 8$). In contrast, leptin (1–20 nM) present in the electrode solution (but not the bath solution) resulted in the activation of potassium channels in approximately 50% of patches (15 of 30 for acutely isolated neurons and 2 of 5 for slice recordings). In all recordings from neurons exposed to 3 mM glucose, in which no channel activity occurred following formation of the cell-attached recording mode, there was a mean delay of 8.7 ± 2.1 min before leptin-induced channel activity was observed (Fig. 3a). For example, N_{TP_0} (average activity—as defined in Methods) increased from 0.0 during the first 1–2 min of recording to 0.41 ± 0.08 ($n = 9$) after 10–20 min exposure to 10 nM leptin. Control experiments, with no leptin in the patch electrode solution, displayed no channel activity over a 30-min period ($n = 6$). The leptin-induced channel activity was inhibited by application of 200 μ M tolbutamide to the bathing solution (Fig. 3b) with a mean inhibition of 56 ± 21.7 ($n = 3$, $P < 0.05$). This sensitivity to tolbutamide suggests that the target channel is K_{ATP} (ref. 13). This was confirmed by examination of inside-out patches formed following leptin (1 nM)-induced activation of single channels in the cell-attached recording configuration

(Fig. 4a). Application of 2 mM ATP resulted in significant ($P < 0.01$), reversible inhibition of channel activity. The mean N_{TP_0} was 0.77 ± 0.31 before ATP application, 0.41 ± 0.27 in the presence of 2 mM ATP and 0.60 ± 0.29 ($n = 4$) following wash. The mean inhibition induced by 2 mM ATP was $60 \pm 12\%$, close to that previously reported for K_{ATP} channels from rat GR neurons^{12,13}.

In a further series of experiments using outside-out membrane patches, we found that leptin activates hypothalamic neurone K_{ATP} channels in a membrane-delimited manner. In one set of experiments we used 1 mM Ca^{2+} in the electrode solution to determine whether BK_{Ca} channels (which have a similar single-channel conductance to K_{ATP} in GR neurons¹⁸) were present, and these patches could therefore be discarded. Outside-out patches, where single-channel activity was low and voltage independent and the currents were of the expected amplitude for K_{ATP} under asymmetrical ionic conditions, were examined for leptin action. In four of seven patches, 1 nM leptin induced a significant ($P < 0.02$) increase in channel activity from an N_{TP_0} value of 0.18 ± 0.16 before leptin exposure to 0.75 ± 0.29 after 7.5 ± 2.8 min. In a second set of experiments ($n = 4$), where the Ca^{2+} concentration in the electrode solution was 100 nM, leptin (1 nM) increased N_{TP_0} from 0.61 ± 0.40

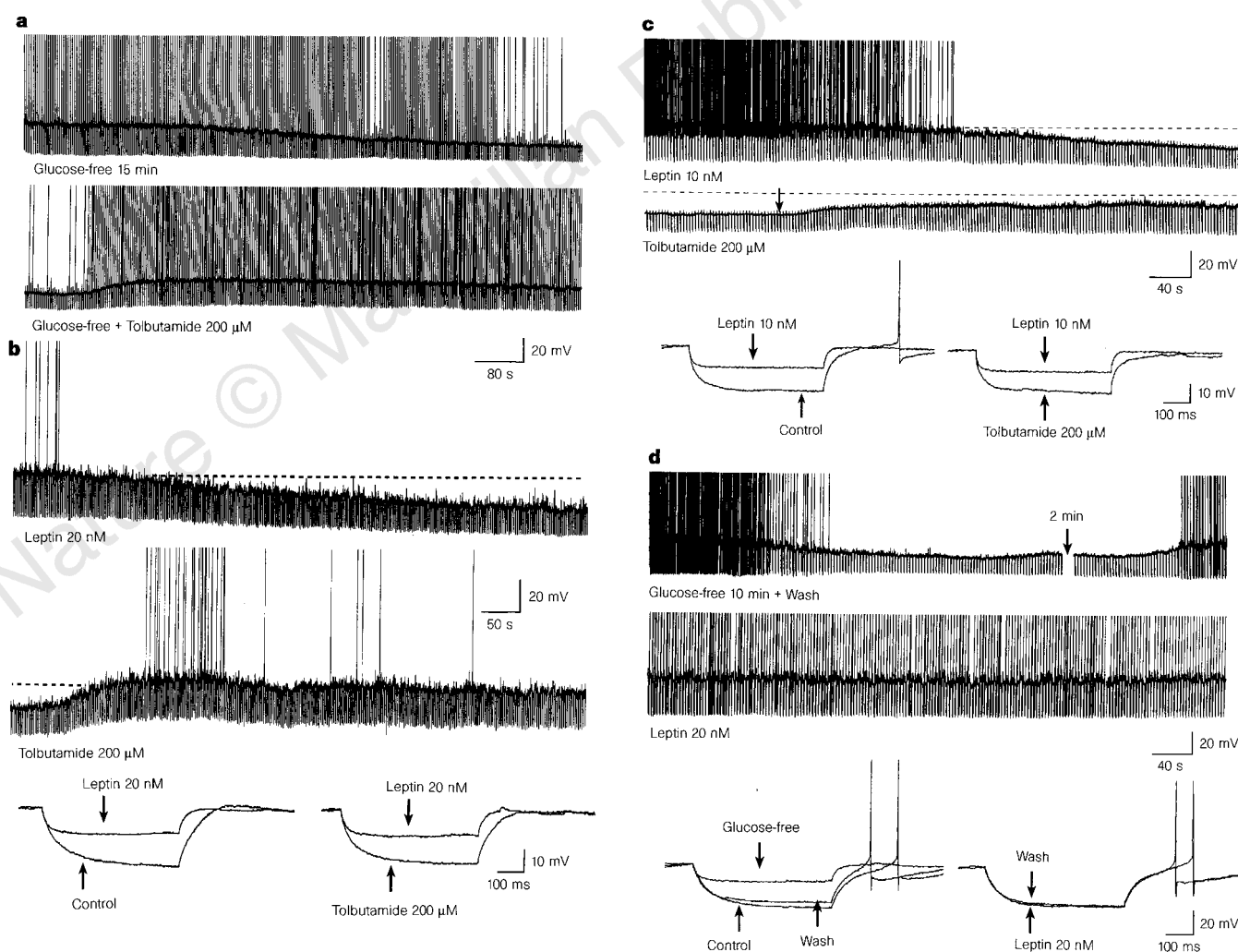


Figure 2 Leptin increases K^+ conductance in GR neurons of Sprague-Dawley, Zucker lean but not obese Zucker rats. **a**, Current-clamp recording showing that reducing glucose concentration from 10 to 0 mM induced hyperpolarization and decreased input resistance. Subsequent application of 200 μ M tolbutamide reversed these actions. **b**, **c**, Recordings from Sprague-Dawley (**b**) and Zucker lean (**c**) rats illustrating leptin-induced (20 and 10 nM, respectively) hyper-

polarization, cessation of action potentials and decreased input resistance; these effects were reversed by 200 μ M tolbutamide. Examples of electrotonic potentials in control aCSF, 20 nM leptin and 200 μ M tolbutamide are shown. **d**, Recording from an obese Zucker rat GR neuron. Leptin (20 nM) had no effect on the neuron membrane properties.

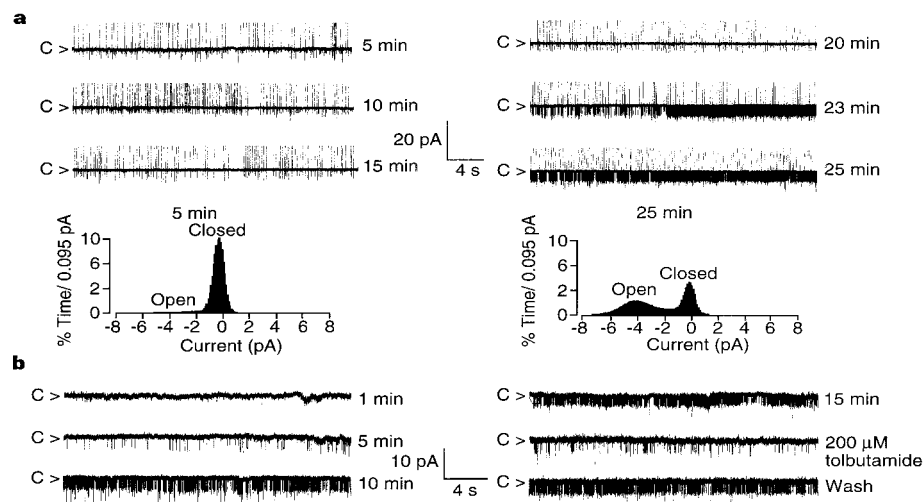


Figure 3 Activation of single, tolbutamide-sensitive K^+ channels by leptin. **a**, Sample traces at various times (indicated on the right) after the beginning of cell-attached recording with 10 nM leptin in the electrode solution. Initially no channel openings occur, but after 20–23 min channel activity (downward deflections) is observed. The biphasic deflections denote action potentials. Their increased frequency in the later traces are due to channel openings and entry of potassium causing depolarization. Frequency histograms denoting mean activity at times indicated are shown underneath. **b**, Cell-attached recording of leptin-induced channel opening and subsequent inhibition by 200 μ M tolbutamide. C denotes the closed state.

to 1.10 ± 0.63 ($P < 0.05$) after approximately 5 min with no change in channel amplitude (Fig. 4b). Channel activity was not inhibited by tolbutamide (200 μ M), in agreement with previous studies of the K_{ATP} channel in isolated patches from GR neurons¹³. The channel was characterized by a linear current–voltage relation, in symmetrical 140 mM KCl, with a slope conductance of 149 ± 25 pS, which is consistent with K_{ATP} of GR neurons^{12,13}.

Our findings that GR neurons of ventromedial and arcuate nuclei are sensitive to leptin at concentrations reported for circulating leptin in animals and humans^{3,19,20} are consistent with these regions of the hypothalamus being the target for leptin action²¹. *In situ* hybridization analysis^{22,23} indicates that Ob- R_L is strongly expressed in the hypothalamus, particularly in arcuate and ventromedial nuclei, and peripheral administration of leptin induces expression of the immediate-early gene *c-fos* in the hypothalamus^{24,25}. Furthermore, the lack of a leptin response in GR neurons from obese Zucker rats strongly indicates that functional receptors are a prerequisite for the hyperpolarizing response in this brain area. This finding is consistent with reports of attenuated leptin effectiveness at reducing food intake and body weight gain when injected into the hypothalamus of obese rats^{26,27}. Recent studies^{28,29} indicate that a similar mechanism may underlie leptin effects on pancreatic β -cells, suggesting metabolic actions in the periphery. As obesity has direct associations with chronic diseases such as hypertension, non-insulin-dependent diabetes and stroke, leptin modulation of K_{ATP} channels, both centrally and peripherally, may form an integral part of the molecular machinery that maintains central and peripheral homeostasis of body weight and energy balance. □

Methods

Recordings from hypothalamic slices. Sprague–Dawley and Zucker rats were anaesthetized and decapitated, the brain was removed, and hypothalamic slices (300 μ m) containing ventromedial and arcuate nuclei were prepared and stored submerged at room temperature until transferral to a submersion-type recording chamber. Slices were continuously superfused with artificial cerebrospinal fluid (aCSF) of the following composition (in mM): NaCl 127, KCl 1.9, KH_2PO_4 1.2, $CaCl_2$ 2.4, $MgCl_2$ 1.3, $NaHCO_3$ 26, D-glucose 10, equilibrated with 95% O_2 , 5% CO_2 , pH 7.4. Whole-cell current and voltage-clamp recordings were obtained, using an Axopatch 1D amplifier (Axon Instruments) at room temperature with patch pipettes (3–10 M Ω) filled with an intracellular solution of the following composition (in mM): K-gluconate 130, KCl 10, $MgCl_2$ 0.1, $CaCl_2$ 0.05, EGTA- Na_3 0.5, HEPES 10, Na_2ATP 2, Lucifer Yellow 2 or biocytin 10, pH 7.4 with NaOH. Input resistance was monitored by the amplitude of electronic potentials, generated by injection of hyperpolarizing rectangular-wave current pulses (–20 to –40 pA, 500 ms, 0.3 Hz). Data were displayed on a chart recorder (Gould Easygraph), and stored on DAT for off-line analysis. Intracellular recordings were obtained at $36 \pm 0.5^\circ C$ with 3 M K-acetate-filled microelectrodes (120–200 M Ω). Recordings were made with an Axoclamp 2A amplifier, and signals monitored as outlined above. Human recombinant leptin (provided by P. Lind, Pharmacia-Upjohn, Stockholm) was prepared as a 1- or 10- μ M stock solution and diluted daily to the concentrations required (4–50 nM) in aCSF containing 2% BSA. Tolbutamide (Sigma) was dissolved in DMSO to give a stock of 100 mM and diluted to a final concentration of 50–200 μ M in aCSF. Drugs were applied by superfusion of the bath with a gravity feed system. Results are expressed as mean $\pm$ s.d.

Recordings from dissociated neurons. Acutely dissociated neurons were

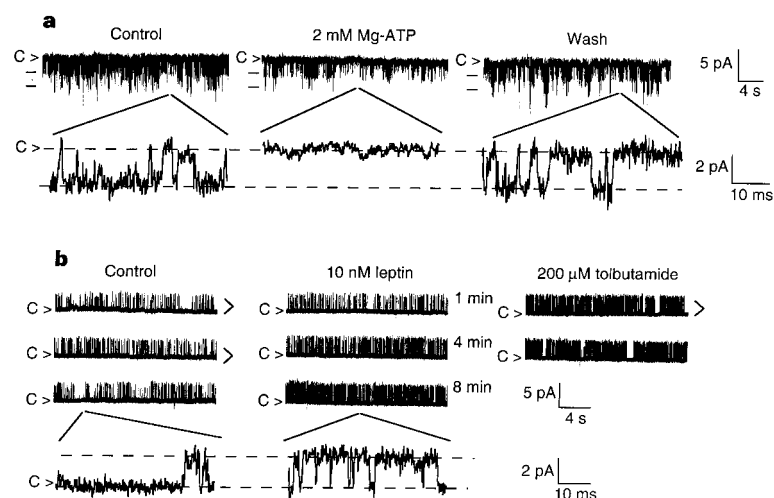


Figure 4 Leptin activates K_{ATP} channels in a membrane-delimited manner. **a**, After cell attached recording and activation of potassium channels by 10 nM leptin, excision into the inside-out configuration resulted in sustained channel activity which was reversibly inhibited by 2 mM Mg-ATP. **b**, Outside-out patch recording under asymmetrical ionic conditions at a potential of 0 mV. Control recordings are shown before leptin addition. Application of 10 nM leptin to the bath solution increased channel activity with time. Tolbutamide (200 μ M) had no effect on channel currents under this recording configuration. The > indicates continuity of recording. C denotes the closed state. Sections of recordings expanded in time show channel activity in more detail.

obtained by dissecting hypothalamic wedges (containing ventromedial and arcuate nuclei) from 350- μ m hypothalamic slices and incubating in 10 ml of 1 mg ml⁻¹ protease (Type XIV, Sigma) in aCSF equilibrated with 95% O₂, 5% CO₂, at room temperature for 1 h, followed by washing 5 times with 30 ml aCSF. The wedges were resuspended in aCSF (10 ml) and dissociated by gentle trituration. Dissociated neurons were plated into tissue-culture dishes containing 1 ml of normal saline and 10 mM glucose and left for 20 min before patching. Recordings were made using an Axopatch 200A patch-clamp amplifier at room temperature with patch-pipettes (5–10 M Ω) filled with the following solution (in mM): KCl 140, CaCl₂ 1, MgCl₂ 1, HEPES 10, pH 7.2; and for some outside-out recordings, KCl 140, CaCl₂ 3.25, MgCl₂ 1, EGTA 10, Mg-ATP 2, HEPES 10, pH 7.2 (free Ca²⁺ of 100 nM). The bath solution (normal saline) consisted of (in mM): NaCl 135, KCl 5, CaCl₂ 1, MgCl₂ 1, HEPES 10, Glucose 3, pH 7.4 with NaOH; for current–voltage relations during inside-out recordings, KCl 140, CaCl₂ 9.75, MgCl₂ 1, EGTA 10, HEPES 10, pH 7.2 (free Ca²⁺ of 10 μ M). Data were displayed on a chart recorder (Gould Easygraph), and stored on DAT for off-line analysis. The potential across the membrane is described following the usual sign convention for membrane potential (inside negative). Outward current is shown as upward deflections on all single-channel records. The data were analysed for current amplitude and average activity (N/P_0) as described previously³⁰. ATP (Sigma) was made up as a 100-mM stock solution in 10 mM HEPES, pH 7.2, and kept at –4 °C until required. Tolbutamide and leptin were made up and applied to the bath as described above. Significance levels in all cases were determined by Student's *t*-test. Results are expressed as the mean $\pm$ s.e.m.

Received 2 June; accepted 15 September 1997.

- Caro, J. F., Sinha, M. K., Kolaczynski, J. W., Zhang, P. L. & Considine, R. V. Leptin: The tale of an obesity gene. *Diabetes* **45**, 1455–1462 (1996).
- Matson, C. A., Wiater, M. F. & Weigle, D. S. Leptin and the regulation of body adiposity. *Diabetes Rev.* **4**, 488–508 (1996).
- Stephens, T. W. *et al.* The role of neuropeptide Y in the antiobesity action of the *obese* gene product. *Nature* **377**, 530–532 (1995).
- Satoh, N. *et al.* The arcuate nucleus as a primary site of satiety effect of leptin in rats. *Neurosci. Lett.* **224**, 149–152 (1997).
- Vaisse, C. *et al.* Leptin activation of Stat3 in the hypothalamus of wild-type and *ob/ob* mice but not *db/db* mice. *Nature Genet.* **14**, 95–97 (1996).
- Levin, B. E. & Routh, V. H. Role of the brain in energy balance and obesity. *Am. J. Physiol.* **40**, R491–R500 (1996).
- Zhang, Y. R. *et al.* Positional cloning of the mouse *obese* gene and its human homologue. *Nature* **372**, 425–432 (1994).
- Lee, G.-H. *et al.* Abnormal splicing of the leptin receptor in *diabetic* mice. *Nature* **379**, 632–635 (1996).
- Chen, H. *et al.* Evidence that the diabetes gene encodes the leptin receptor: Identification of a mutation in the leptin receptor gene in *db/db* mice. *Cell* **84**, 491–495 (1996).
- Chua, S. C. Jr *et al.* Phenotypes of mouse *diabetes* and rat *fatty* due to mutations in the OB (leptin) receptor. *Science* **271**, 994–996 (1996).
- Oomura, Y., Ono, T., Ooyama, H. & Wayner, M. J. Glucose and osmosensitive neurones in the rat hypothalamus. *Nature* **222**, 282–284 (1969).
- Ashford, M. L. J., Boden, P. R. & Treherne, J. M. Glucose-induced excitation of hypothalamic neurones is mediated by ATP-sensitive K⁺ channels. *Pflügers Arch.* **415**, 479–483 (1990).
- Ashford, M. L. J., Boden, P. R. & Treherne, J. M. Tolbutamide excites rat glucoreceptive ventromedial hypothalamic neurones by indirect inhibition of ATP-K⁺ channels. *Br. J. Pharmacol.* **101**, 531–540 (1990).
- Streamson, C. *et al.* Phenotype of *fatty* due to Gln269Pro mutation in the leptin receptor (*Lepr*). *Diabetes* **45**, 1141–1143 (1996).
- Phillips, M. S. *et al.* Leptin receptor missense mutation in the *fatty* Zucker rat. *Nature Genet.* **13**, 18–19 (1996).
- Rosenblum, C. I. *et al.* Functional STAT 1 and 3 signalling by the leptin receptor (OB-R); reduced expression of the fatty leptin receptor in transfected cells. *Endocrinology* **137**, 5178–5181 (1996).
- Yamashita, T., Murakami, T., Iida, M., Kuwajima, M. & Shima, K. Leptin receptor of Zucker fatty rat performs reduced signal transduction. *Diabetes* **46**, 1077–1080 (1997).
- Treherne, J. M. & Ashford, M. L. J. Calcium-activated potassium channels in rat dissociated ventromedial hypothalamic neurones. *J. Neuroendocrinol.* **3**, 323–329 (1991).
- Tartaglia, L. A. *et al.* Identification and expression cloning of a leptin receptor, OB-R. *Cell* **83**, 1263–1271 (1995).
- Rock, F. L., Peterson, D., Weig, B. C., Kastelein, R. A. & Bazan, J. F. Binding of leptin to the soluble ectodomain of recombinant leptin receptor. *Horm. Metab. Res.* **28**, 748–750 (1996).
- Glaum, S. R. *et al.* Leptin, the *Obese* gene product, rapidly modulates synaptic transmission in the hypothalamus. *Mol. Pharmacol.* **50**, 230–235 (1996).
- Mercer, J. G. *et al.* Localization of leptin receptor mRNA and the long form splice variant (Ob-Rb) in mouse hypothalamus and adjacent brain regions by *in situ* hybridization. *FEBS Lett.* **387**, 113–116 (1996).
- Schwartz, M. W., Seeley, R. J., Campfield, L. A., Burn, P. & Baskin, D. G. Identification of targets of leptin action in rat hypothalamus. *J. Clin. Invest.* **98**, 1101–1106 (1996).
- Woods, A. J. & Stock, M. J. Leptin activation in hypothalamus. *Nature* **381**, 745 (1996).
- Elmqvist, J. K., Ahima, R. S., Maratos-Flier, E., Flier, J. S. & Saper, C. B. Leptin activates neurons in ventrobasal hypothalamus and brainstem. *Endocrinology* **138**, 839–842 (1997).
- Cusin, I., Rohner-Jaenrenaud, E., Sticker-Krongrad, A. & Jaenrenaud, B. The weight-reducing effect of an intracerebroventricular bolus injection of leptin in genetically obese *fa/fa* rats: reduced sensitivity compared with lean animals. *Diabetes* **45**, 1446–1450 (1996).
- Seeley, R. J. *et al.* Intraventricular leptin reduces food intake and body weight of lean rats but not obese Zucker rats. *Horm. Metab. Res.* **28**, 664–668 (1996).

- Keiffer, T. J., Heller, R. S., Leech, C. A., Holz, G. G. & Habener, J. F. Leptin suppression of insulin secretion by the activation of ATP-sensitive K⁺ channels in pancreatic β -cells. *Diabetes* **26**, 1087–1093 (1997).
- Harvey, J., McKenna, F., Herson, P. S., Spanswick, D. & Ashford, M. L. J. Leptin activates ATP-sensitive potassium channels in the rat insulin secreting cell line, CRI-G1. *J. Physiol. (Lond.)* **504**, 527–535 (1997).
- Lee, K., Rowe, I. C. M. & Ashford, M. L. J. Characterization of an ATP-modulated large conductance Ca²⁺-activated K⁺ channel present in rat cortical neurones. *J. Physiol. (Lond.)* **488**, 319–337 (1995).

Acknowledgements. We thank K. Todd for technical assistance. This work was supported by grants from the Wellcome Trust and Pharmacia-Upjohn.

Correspondence and requests for materials should be addressed to M.L.J.A. (e-mail: mike.ashford@abdn.ac.uk).

RGS8 accelerates G-protein-mediated modulation of K⁺ currents

Osamu Saitoh*, Yoshihiro Kubo†, Yoshihide Miyatani*, Tomiko Asano‡ & Hiroyasu Nakata*

* Department of Molecular and Cellular Neurobiology, † Department of Neurophysiology, Tokyo Metropolitan Institute for Neuroscience, 2-6 Musashidai, Fuchu-shi, Tokyo 183, Japan

‡ Department of Biochemistry, Institute for Developmental Research, Aichi Human Service Center, Kasugai, Aichi 480-03, Japan

Transmembrane signal transduction via heterotrimeric G proteins is reported to be inhibited by RGS (regulators of G-protein signalling) proteins^{1–4}. These RGS proteins work by increasing the GTPase activity of G protein α -subunits (G α), thereby driving G proteins into their inactive GDP-bound form^{5–7}. However, it is not known how RGS proteins regulate the kinetics of physiological responses that depend on G proteins. Here we report the isolation of a full-length complementary DNA encoding a neural-tissue-specific RGS protein, RGS8, and the determination of its function. We show that RGS8 binds preferentially to the α -subunits G α_o and G α_i3 and that it functions as a GTPase-activating protein (GAP). When co-expressed in *Xenopus* oocytes with a G-protein-coupled receptor and a G-protein-coupled inwardly rectifying K⁺ channel (GIRK1/2), RGS8 accelerated not only the turning off but also the turning on of the GIRK1/2 current upon receptor stimulation, without affecting the dose–response relationship. We conclude that RGS8 accelerates the modulation of G-protein-coupled channels and is not just a simple negative regulator. This property of RGS8 may be crucial for the rapid regulation of neuronal excitability upon stimulation of G-protein-coupled receptors.

To identify RGS proteins specifically expressed in neural cells, we used the polymerase chain reaction (PCR) on cDNA from neuronally differentiated P19 mouse embryonal carcinoma cells⁸, with degenerate oligonucleotide primers corresponding to the conserved region (RGS domains) of known RGS proteins³. Three genes that encode RGS domains were isolated and identified as *RGS-r* (ref. 9), *RGS2* (refs 4, 10) and *RGS8* (ref. 3), respectively. We next investigated messenger RNA expression for these three RGS proteins in undifferentiated and neuronally differentiated P19 cells by reverse transcription-PCR (RT-PCR) analysis. Although expression of *RGS2* and *RGS-r* mRNA was similar in both types of cell, that of *RGS8* mRNA was significantly higher in differentiated cells (Fig. 1a).

As only a partial cDNA fragment of *RGS8* has been isolated³, we screened a rat hippocampus cDNA library with a PCR-amplified fragment of *RGS8* in order to isolate a full-length cDNA clone for functional analysis. The isolated cDNA encoded a protein of 180 amino acids (Fig. 1b). Comparison of the predicted full-length amino-acid sequence with the GenBank data base revealed significantly similarity with several members of the RGS family. The predicted amino-acid sequence was identical to the partial sequence reported for *RGS8* (ref. 3). Northern blot analysis indicated that

RGS8 was expressed at high levels in rat brain and barely at all in other tissues (Fig. 1c). A smaller transcript was detected in testis by the RGS8 probe (Fig. 1c, lane 8). During brain development, expression of RGS8 was detectable in 13-day-old embryos, increasing gradually in later embryos and then markedly in neonates to adults (Fig. 1d).

We next determined the specificity of RGS8 by *in vitro* binding⁶. Hexa-histidine tagged RGS8 (His-tagged RGS8) was incubated with rat brain membrane fractions treated with GDP and AlF_4^- , as other RGS proteins are known to have stronger binding to G proteins in the presence of GDP and AlF_4^- (ref. 6). Complexes containing the His-tagged RGS8 were purified on Ni^{2+} -NTA-agarose and were analysed by SDS-polyacrylamide gel electrophoresis (SDS-PAGE). After Coomassie blue staining, a protein of relative molecular mass 40,000 (M_r 40K) recovered with RGS8 (Fig. 2a). By immunoblotting using antisera specific for different subtypes of $G\alpha$ subunit, we found that $G\alpha_o$ and $G\alpha_{i3}$ were selectively recovered with RGS8. Only weak signals were visible for $G\alpha_i$ 1 and 2, $G\alpha_z$ and $G\alpha_{q/11}$ (Fig. 2b). Reconstitution studies have shown that RGS proteins (RGS1, 4, 10 and GAIP) interact selectively with the Gi class of $G\alpha$ subunits but that they do not discriminate individual $G\alpha_i$ sub-

types⁵⁻⁷. Our results show that RGS8 interacts selectively with $G\alpha_o$ and $G\alpha_{i3}$ in brain membranes where various subtypes of $G\alpha$ exist.

To investigate how RGS8 regulates the function of $G\alpha$ subunits, we tested the effects of RGS8 on the GTP hydrolysis rate of $G\alpha_o$ and on the GDP-GTP exchange rate of $G\alpha_o$. The GTP hydrolysis rate was measured by first making GTP bind to $G\alpha_o$ under conditions in which the nucleotide cannot be hydrolysed, so that the GDP-GTP exchange did not limit the rate. RGS8 markedly increased the rate of GTP hydrolysis of bovine $G\alpha_o$, whereas boiled RGS8 had no effect (Fig. 3a). These results indicate that RGS8 functions as a GAP for $G\alpha_o$. The GDP-GTP exchange rate of $G\alpha_o$ was estimated by measuring the steady-state GTP hydrolysis rate, which is limited not by the rate of catalysis but by the slow exchange of GDP and GTP. RGS8 did not significantly affect the GDP-GTP exchange rate under these conditions (Fig. 3b). However, as the estimation of GDP-GTP exchange rate here was done in the absence of stimulation by an activated receptor, it remains to be seen whether RGS8 alters the fast GDP-GTP exchange rate on $G\alpha_o$ upon stimulation by receptor. As this step is probably too fast to be analysed biochemically, we decided to examine the kinetics of the response upon receptor activation by using electrophysiological recording.

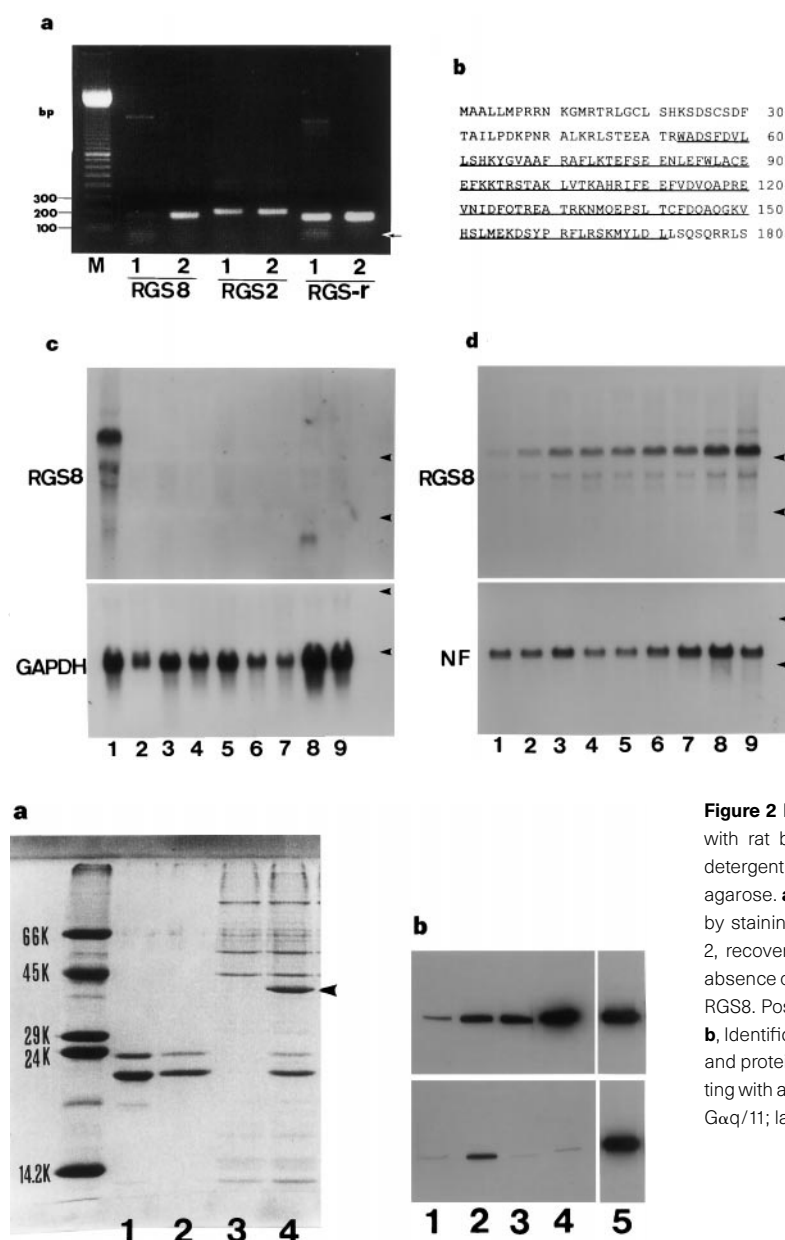


Figure 1 Detection and characterization of RGS8. **a**, Expression of RGS8, RGS2 and RGS-r during neural differentiation of P19 cells. Expression of RGS8, RGS2 and RGS-r in undifferentiated (lane 1) and neuronally differentiated (lane 2) P19 cells was examined by using RT-PCR; the position of the primers is indicated by an arrow. DNA size markers (100-bp ladder) are also shown (M). The expected size of the PCR product is ~200 bp. **b**, Amino-acid sequences of rat RGS8; the RGS domain is underlined. **c**, Northern blot analysis of various rat tissues. Lanes: 1, brain; 2, heart; 3, lung; 4, stomach; 5, spleen; 6, liver; 7, kidney; 8, testis; and 9, back muscle (M. latissimus dorsi). Total RNA (20 μ g) was electrophoresed, transferred and then hybridized with full-length rat RGS8 cRNA (top) or mouse glyceraldehyde 3-phosphate dehydrogenase (GAPDH) cDNA (bottom). **d**, Northern blot analysis of RNA from developing brain. Lanes: 1, heads of 13-d embryos; 2-6, brains from 14-d, 15-d, 17-d, 19-d, 21-d embryos, respectively; 7, 8, brains from 6-d, 13-d neonates; and 9, adults. Total RNA (10 μ g) was electrophoresed, transferred and then hybridized with full-length rat RGS8 cRNA (top) or mouse neurofilament (NF140K) cDNA (bottom). Probes for GAPDH and NF140K have been described²⁴. Arrowheads, migration positions of 28S and 18S rRNA markers.

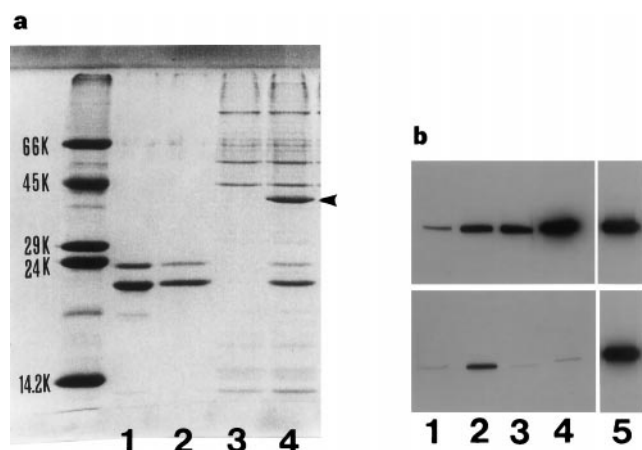


Figure 2 Interaction of RGS8 with $G\alpha$ subunit. His-tagged RGS8 was incubated with rat brain membranes treated with GDP and AlF_4^- . After extraction with detergent, complexes containing His-tagged RGS8 were purified by Ni^{2+} -NTA-agarose. **a**, Proteins bound to RGS8 were resolved by SDS-PAGE and detected by staining with Coomassie brilliant blue. Lanes: 1, purified His-tagged RGS8; 2, recovered His-tagged RGS8 from buffer alone; 3, proteins recovered in the absence of His-tagged RGS8; 4, proteins recovered in the presence of His-tagged RGS8. Position of the 40K protein recovered with His-tagged RGS8 is indicated. **b**, Identification of $G\alpha$ subunit bound to RGS8. Rat-brain membrane proteins (top) and proteins bound to His-tagged RGS8 (bottom) were analysed by immunoblotting with antibodies against: lane 1, $G\alpha_i$ 1 and $G\alpha_{i2}$; lane 2, $G\alpha_{i3}$; lane 3, $G\alpha_z$; lane 4, $G\alpha_{q/11}$; lane 5, $G\alpha_o$.

G-protein-coupled inwardly rectifying K^+ channels are activated directly by $G\beta\gamma$ subunits¹¹ released from pertussis-toxin-sensitive G proteins of the G_i family, including G_i and G_o (ref. 12). They are activated by various G-protein-coupled receptors such as the m2 muscarinic receptor¹³, the D2 dopamine receptor¹³, metabotropic glutamate receptors¹⁴, the GABA_B receptor¹⁵, and μ and δ opioid receptors¹⁶. As turning off the G-protein-coupled K^+ channel upon removal of the agonist is much faster than the rate of basal GTP hydrolysis by purified $G\alpha$ subunits¹⁷, soluble factors like GAP have been proposed to be responsible for this rapid switching off¹⁸.

We co-expressed the brain-type G-protein-coupled inwardly rectifying K^+ channels, GIRK1¹⁹/GIRK2 heteromultimer^{16,20,21} and D2 dopamine receptor (Fig. 4a) or m2 muscarinic receptor (Fig. 4b) with or without RGS8 in *Xenopus* oocytes, and analysed the speed of turning on and off upon agonist application under two-electrode voltage clamp. Co-expression of RGS8 accelerated both turning on and off (Fig. 4a, b). The increasing and decreasing phases were well fitted by a single exponential function, and the time constants for the fit, τ_{on} and τ_{off} are shown in Fig. 4c. The acceleration of turning on and off by RGS8 were statistically significant. Results were similar when cardiac-type G-protein-coupled inwardly rectifying K^+ channel GIRK1/CIR²² was used instead of the brain type GIRK1/GIRK2 (data not shown).

How can the acceleration of both the turning on and off of the GIRK current be explained? One possibility is that RGS8 has a dual function: that is, it can act as an off-rate accelerator (GAP) but also as an on-rate accelerator (for example, an accelerator of GDP–GTP exchange or of dissociation of $G\alpha$ and $G\beta\gamma$). Another possibility is that RGS8 simply functions as a GAP, and that the time required to reach equilibrium upon receptor activation ($1/(\text{on-rate} + \text{off-}$

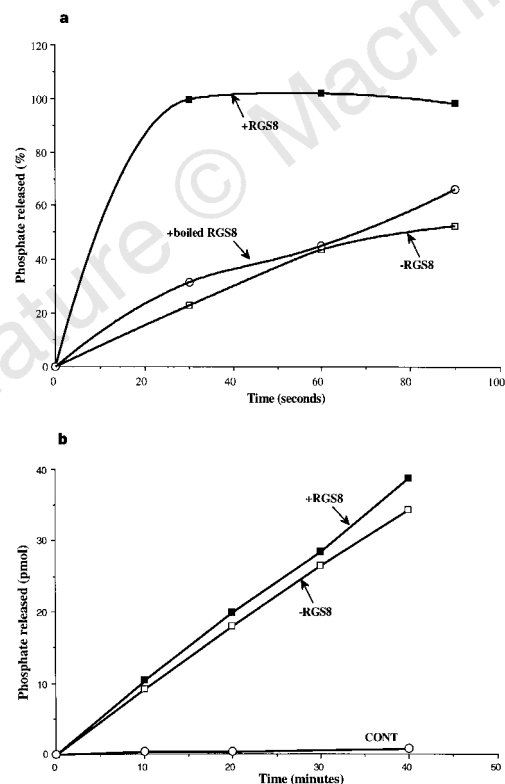


Figure 3 Effects of RGS8 on GTP hydrolysis of $G\alpha_o$. **a**, The kinetics of GTP hydrolysis during a single catalytic turnover by $G\alpha_o$ alone (open squares) or by $G\alpha_o$ with RGS8 (filled squares) or with boiled RGS8 (circles). Released phosphate is expressed as a percentage of the maximum level (3.3 pmol). **b**, Kinetics of steady-state hydrolysis of GTP by $G\alpha_o$ alone (open squares) or $G\alpha_o$ with RGS8 (filled squares) or buffer (circles). Data are representative of at least two independent experiments.

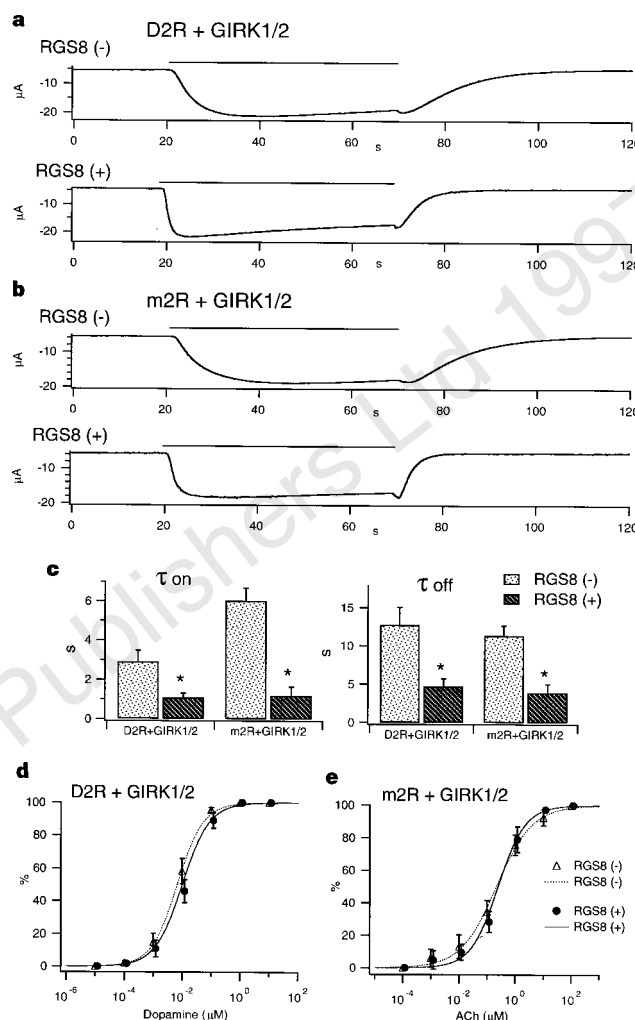


Figure 4 Effects of RGS8 on turning-on and turning-off kinetics, and on the dose-response relationships of the G-protein-coupled inwardly rectifying K^+ channel (GIRK1/2) current upon stimulation of the receptors. **a**, GIRK1/2 and D2 dopamine receptor without (upper trace) or with (lower trace) RGS8 were co-expressed in *Xenopus* oocytes. Current traces at a holding potential of -120 mV are shown. 50 nM dopamine was applied at the time indicated by the bars. The kinetics of the turning on and off are faster for co-expression with RGS8. **b**, Same as in **a**, except that co-expression was with the m2 muscarinic receptor instead of D2 dopamine receptor. ACh (4 μ M) was applied at the time indicated by the bars. **c**, Comparison of the time constants τ_{on} and τ_{off} of the GIRK1/2 current upon stimulation of the D2 or m2 receptor, in the absence or presence of RGS8. From the traces in **a** and **b**, the increasing and decreasing phases were fitted by a single exponential function and τ_{on} and τ_{off} were obtained. The mean and standard deviation of τ_{on} ($n = 14$) were: (D2 receptor, RGS(-)); $2,905 \pm 576$ ms; (D2 receptor, RGS(+)); $1,103 \pm 206$ ms; (m2 receptor, RGS(-)); $6,021 \pm 651$ ms; (m2 receptor, RGS(+)); $1,188 \pm 444$ ms. Those of τ_{off} ($n = 14$) were: (D2 receptor, RGS(-)); $12,761 \pm 2,360$ ms; (D2 receptor, RGS(+)); $4,714 \pm 998$ ms; (m2 receptor, RGS(-)); $11,361 \pm 1,304$ ms; (m2 receptor, RGS(+)); $3,909 \pm 1,087$ ms. Asterisks indicate data judged to be significantly different from control values (P value < 0.05) by Student's unpaired t -test. **d**, Comparison of dose-response relationships upon stimulation of the D2 receptor in the absence (triangles, broken line) and presence (filled circles, solid line) of RGS8. Responses to all doses were tested in a single oocyte and normalized. The average and s.d. of five oocytes were plotted. K_d values and Hill coefficients for fitting were 6.7 nM, 0.96 , respectively, for RGS8(-), and 10 nM, 0.92 for RGS8(+). **e**, Same as in **d**, except that the m2 instead of D2 receptor was used. K_d values and Hill coefficients for fitting were 0.21 μ M, 0.68 for RGS8(-), and 0.25 μ M, 0.89 for RGS8(+).

rate)) is shortened by the increase in off-rate. If so, the equilibrium should be shifted towards the GDP-bound form. However, the equilibrium was not shifted, because the dose-response relationships (Fig. 4d, e) and the amplitudes of basal and induced currents (data not shown) were not significantly changed by RGS8. Thus our observations cannot all be explained by assuming that RGS8 functions only as a GAP. We therefore propose that RGS8 functions not only as a GAP but also as an on-rate accelerator, for example in promoting GDP-GTP exchange or the dissociation of $G\alpha$ and $G\beta\gamma$.

We have isolated a full-length cDNA encoding RGS8, a unique member of the RGS family. Specific features of RGS8 are its expression in the brain and its preferred binding to $G\alpha_o$ and $G\alpha_{i3}$. Biochemical reconstitution has revealed that RGS8 acts as a GAP for $G\alpha$ and functional studies in *Xenopus* oocytes showed that RGS8 significantly accelerates the turning on of the K^+ channel upon addition of agonist, as well as turning off when the agonist is washed out. Although RGS proteins have been proposed to be negative regulators of G-protein-coupled signal transduction¹, our results indicate that they also accelerate the turning-on and -off kinetics of G-protein signalling. As RGS8, $G\alpha_o$ and GIRK1/2 are all expressed in the brain, this accelerated response caused by RGS8 may occur in the brain to enable the resting potential and the firing of neuronal cells to be rapidly regulated.

Note added in proof: Doupnik *et al.* reported similar electrophysiological results on RGS 1, 3, and 4 (*Proc. Natl Acad. Sci. USA* **94**, 10461–10466; 1997). □

Methods

RT-PCR. To identify RGS proteins expressed in neuronally differentiated P19 cells, RT-PCR was used with degenerate primers corresponding to conserved RGS domains as described³. Total RNA was isolated from P19 cells differentiated by retinoic acid treatment. The first strand cDNA was synthesized as a template of PCR. The amplified 200-bp DNA was cloned into pGEM-T vector (Promega) and its sequence determined. Three rat RGS proteins, RGS8, RGS2 and RGS-r were identified. Primers specific for these three RGS proteins (RGS8: 5'-AGTTCAGAAGACCAGGTCAACT-3', 5'-TTTTCCTGAGCTTGATCAAAACAA-3'. RGS2: 5'-TTGGCTTGTAAGACTTCAA-3', 5'-AAAC-TGTACACCTCTTCTGA-3', RGS-r: AGTTCAGAAGATCCGATCAGCCA-3', 5'-ACATCGAAGCAACTGGTAGT-3') were synthesized and the expression of each RGS was examined by RT-PCR using RNAs isolated from undifferentiated and differentiated P19 cells.

Cloning of cDNA. Rat hippocampus cDNA library (generously provided by K. Yamagata) was screened with the PCR-amplified fragment of RGS8. The longest cDNA clone was sequenced on both strands.

Expression and purification of His-tagged RGS8. RGS8 was expressed as a full-length protein with hexa-histidine tags at the amino terminus. A bacterial expression plasmid of His-tagged RGS8 was constructed as follows. The entire coding region of RGS8 cDNA was used to generate a PCR fragment and the nucleotide sequence of the amplified DNA was confirmed by sequencing both strands. The DNA fragment encoding RGS8 was cloned into the *Bam*HI and *Sma*I sites of pQE30 (Qiagen). The resultant plasmid was transformed into *E. coli*, strain M15. His-tagged RGS8 was induced with 2 mM IPTG for 4 h and extracted by sonication in 50 mM sodium phosphate, pH 7.4, 1.3 M NaCl, 10% glycerol and 40 mM imidazole (sonication buffer). Lysates were clarified by centrifugation and then applied to Ni^{2+} -NTA-agarose (Qiagen) columns. Columns were washed with sonication buffer containing 10 mM β -mercaptoethanol. The proteins were eluted in a stepwise manner with increasing concentrations of imidazole. His-tagged RGS8 was eluted with 0.4 M imidazole.

Binding assay of RGS8. RGS8-G protein binding was assayed as described⁶. His-tagged RGS8 (10 μ g) was incubated for 30 min at 5 °C with rat brain membrane fractions (0.5 mg) treated with 10 μ M GDP and 30 μ M AlF_4 . After extraction with 1% cholate, complexes containing His-tagged RGS8 isolated with Ni^{2+} -NTA-agarose were examined by SDS-PAGE. Proteins bound to RGS8 were identified by immunoblotting with antibodies against $G\alpha_{i1}$ and $G\alpha_{i2}$, $G\alpha_{i3}$, (Calbiochem-Novabiochem), $G\alpha_o$, $G\alpha_z$ and $G\alpha_q/11$ (Santa Cruz Biotechnology). Signals were detected with an ECL system (Amersham). The membrane filters were exposed to X-ray films (Hyperfilm, Amersham) for

2 min to detect $G\alpha_o$ and for 6 min for the other $G\alpha$ proteins.

Analysis of GTP hydrolysis. $G\alpha_o$ was purified from bovine brain as described²³. GTPase was assayed as described². To obtain the catalytic rate of GTP hydrolysis, $G\alpha_o$ (0.21 μ M) was loaded with 0.55 μ M [γ -³²P]GTP in 50 mM Na-HEPES (pH 8.0), 2 mM dithiothreitol, 5 mM EDTA and 0.05% C12E10 for 15 min at 20 °C, then the temperature was reduced by placing on ice for 5 min. Before initiation of GTP hydrolysis, a 50- μ l aliquot of the mixture was removed and added to 750 μ l of 5% (w/v) Norit in 50 mM NaH_2PO_4 (0 °C) as the zero time point. Then, unlabelled GTP (166 μ M), $MgSO_4$ (16.6 mM), and His-tagged RGS8 (1 μ M) or buffer were added to the reaction mixture. Aliquots (50 μ l) were taken at the indicated intervals and added to 5% Norit. The charcoal was removed by centrifugation and 400- μ l aliquots of supernatant containing ³²P_i were counted by liquid scintillation spectrometry. To examine steady-state hydrolysis of GTP, $G\alpha_o$ alone (0.24 μ M) or $G\alpha_o$ with 0.88 μ M His-tagged RGS8 were incubated at 20 °C with 6 mM $MgSO_4$, 1.27 μ M [γ -³²P]GTP, and 3.2 μ M cold GTP in 50 mM NaHEPES (pH 8.0), 1 mM EDTA, 2 mM dithiothreitol, and 0.05% C12E10. Aliquots (50 μ l) were removed at the indicated times and processed as described.

Two-electrode voltage clamp. *Xenopus* oocytes were treated with collagenase (2 mg ml⁻¹) for 2 h at room temperature to remove follicle cells and injected with 50 nl of *in vitro* transcribed RNA solution (1 μ g μ l⁻¹). Electrophysiological recordings were made 2–4 d later at room temperature (23 $\pm$ 2 °C) under two-electrode voltage clamp (OC-725B-HV, Warner). Data were acquired and analysed on a 80486-based computer using Digidata 1200 and pCLAMP program (Axon Instruments). Intracellular glass microelectrodes were filled with 3 M KCl; the resistance was 0.2–0.8 M ohm. The bath solution contained 90 mM KCl, 3 mM $MgCl_2$, 20 μ M $GdCl_3$, 10 mM HEPES (pH 7.4). Concentrated stock (5 $\times$) agonist solution was applied to the bath (20% of bath volume) so that the flow was not directed at the oocytes, and then mixed immediately by pipetting. Rapid perfusion and vacuum suction was used to wash out agonists in the bath. The exchange rate of solutions in the bath was determined by changing the K^+ concentration using the same methods. The value τ_{in} ranged from 299 to 533 ms, and τ_{out} from 840 to 1,132 ms; this speed is sufficient for the accurate analysis of the off-rate. The fast on-rate in the presence of RGS8 is estimated to be slightly slower than the real value. This error does not affect our conclusion that turning on is faster when RGS8 is co-expressed.

Received 29 July; accepted 5 September 1997.

1. Koelle, M. R. A new family of G-protein regulators—the RGS protein. *Curr. Opin. Cell Biol.* **9**, 143–147 (1997).
2. Dohlgan, H., Apanieski, D., Chen, Y., Song, J. & Dlussek, D. Inhibition of G-protein signaling by dominant gain-of-function mutation in Sst2p, a pheromone desensitization factor in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **15**, 3635–3643 (1995).
3. Koelle, M. R. & Horvitz, H. R. EGL-10 regulates G protein signaling in the *C. elegans* nervous system and shares a conserved domain with many mammalian proteins. *Cell* **84**, 115–125 (1996).
4. Druey, K. M., Blumer, K. J., Kang, V. H. & Kehrl, J. H. Inhibition of G-protein-mediated MAP kinase activation by a new mammalian gene family. *Nature* **379**, 742–746 (1996).
5. Berman, D. M., Wilkie, T. M. & Gilman, A. G. GAI and RGS4 are GTPase-activating proteins for the G_i subfamily of G protein α subunits. *Cell* **86**, 445–452 (1996).
6. Watson, N., Linder, M. E., Druey, K. M., Kehrl, J. H. & Blumer, K. J. RGS family members: GTPase-activating proteins for heterotrimeric G-protein α -subunits. *Nature* **383**, 172–175 (1996).
7. Hunt, T. W., Fields, T. A., Casey, P. J. & Peralta, E. G. RGS10 is a selective activator of G α_{i2} GTPase activity. *Nature* **383**, 175–177 (1996).
8. Jones-Villeneuve, E. M. V., McBurney, M. W., Rogers, K. A. & Kalnins, V. I. Retinoic acid induces embryonal carcinoma cells to differentiate into neurons and glial cells. *J. Cell Biol.* **94**, 253–262 (1982).
9. Chen, C.-K., Wieland, T. & Simon, M. I. RGS-r, a retinal specific RGS protein, binds an intermediate conformation of transducin and enhances recycling. *Proc. Natl Acad. Sci. USA* **93**, 12885–12889 (1996).
10. Siderovski, D. P., Heximer, S. P. & Forsdyke, D. R. A human gene encoding a putative basic helix-loop-helix phosphoprotein whose mRNA increases rapidly in cycloheximide-treated blood mononuclear cells. *DNA Cell Biol.* **13**, 125–147 (1994).
11. Reuveny, E. *et al.* Activation of the cloned muscarinic potassium channel by G protein $\beta\gamma$ subunits. *Nature* **370**, 143–146 (1994).
12. Wickman, K. & Clapham, D. E. Ion channel regulation by G proteins. *Physiol. Rev.* **75**, 865–885 (1995).
13. Werner, P., Hussy, N., Buell, G., Jones, K. A. & North, R. A. D2, D3, and D4 dopamine receptors coupled to G protein-regulated potassium channels in *Xenopus* oocytes. *Mol. Pharmacol.* **49**, 656–661 (1996).
14. Saugstad, J. A., Segerson, T. P. & Westbrook, G. L. Metabotropic glutamate receptors activate G-protein-coupled inwardly rectifying potassium channels in *Xenopus* oocytes. *J. Neurosci.* **16**, 5979–5985 (1996).
15. Soderstrom, D. L. & Bean, B. P. GABA_B receptor-activated inwardly rectifying potassium current in dissociated hippocampal CA3 neurons. *J. Neurosci.* **15**, 6374–6385 (1996).
16. Lesage, F. *et al.* Cloning provides evidence for a family of inward rectifier and G-protein coupled K^+ channels in the brain. *FEBS Lett.* **353**, 37–42 (1994).
17. Breitwieser, G. E. & Szabo, G. Mechanism of muscarinic receptor-induced K^+ channel activation as revealed by hydrolysis-resistant GTP analogues. *J. Gen. Physiol.* **91**, 469–493 (1988).
18. Shui, Z., Boyett, M. R., Zang, W. J., Haga, T. & Kameyama, K. Receptor kinase-dependent desensitization of the muscarinic K^+ current in rat atrial cells. *J. Physiol. (Lond.)* **487**, 359–366 (1995).

19. Kubo, Y., Reuveny, E., Slesinger, P. A., Jan, Y. N. & Jan, L. Y. Primary structure and functional expression of a rat G-protein-coupled muscarinic potassium channel. *Nature* **364**, 802–806 (1993).
20. Lesage, F. *et al.* Molecular properties of neuronal G-protein-activated inwardly rectifying K⁺ channels. *J. Biol. Chem.* **270**, 28660–28667 (1995).
21. Velimirovic, B. M., Gordon, E. A., Lim, N. F., Navarro, B. & Clapham, D. E. The K⁺ channel inward rectifier subunits form a channel similar to neuronal G-protein-gated K⁺ channel. *FEBS Lett.* **379**, 31–37 (1996).
22. Krapivinsky, G. *et al.* The G-protein-gated atrial K⁺ channel I_{KACH} is a heteromultimer of two inwardly rectifying K⁺-channel proteins. *Nature* **374**, 135–141 (1995).
23. Asano, T., Kamiya, N., Morishita, R. & Kato, K. Immunoassay for the $\beta\gamma$ subunits of GTP-binding proteins and their regional distribution in bovine brain. *J. Biochem.* **103**, 950–953 (1988).
24. Abe, H., Saitoh, O., Nakata, H., Yoda, A. & Matsuda, R. Expression of neurofilament proteins in proliferating C2C12 mouse skeletal muscle cells. *Exp. Cell Res.* **229**, 48–59 (1996).

Acknowledgements. We thank M. Lazdunski for GIRK2 cDNA, A. Connolly for m2 muscarinic receptor and D2 dopamine receptor cDNA, K. Yamagata for the cDNA library, M. Odagiri for technical assistance. This work is supported by research grants from the Ministry of Education, Science, Sports and Culture of Japan (to O.S.) and from the Human Frontier Science Program Organization (to Y.K.) and by a grant-in-aid for scientific research on priority area of "channel-Transporter Correlation" (to Y.K.).

Correspondence and requests for materials should be addressed to O.S. (e-mail: osaito@tmin.ac.jp) or H.N. (e-mail: nakata@tmin.ac.jp). The nucleotide sequence of rat RGS8 will appear in the DDBJ, EMBL, and GenBank data bases under the accession number AB006013.

Pore-forming segments in voltage-gated chloride channels

Christoph Fahlke, Henry T. Yu, Carol L. Beck, Thomas H. Rhodes & Alfred L. George Jr

Departments of Medicine (Nephrology) and Pharmacology, and the Center for Molecular Neuroscience, Vanderbilt University School of Medicine, Nashville, Tennessee 37232-2372, USA

The ability to differentiate between ions is a property of ion channels that is crucial for their biological functions¹. However, the fundamental structural features that define anion selectivity and distinguish anion-permeable from cation-permeable channels are poorly understood. Voltage-gated chloride (Cl⁻) channels belonging to the ClC family are ubiquitous and have been predicted to play important roles in many diverse physiological² and pathophysiological^{3–5} processes. We have identified regions of a human skeletal muscle ClC isoform that contribute to formation of its anion-selective conduction pathway. A core structural element (P1 region) of the ClC channel pore spans an accessibility barrier between the internal and external milieu, and contains an evolutionarily conserved sequence motif: GKxGPxxH. Neighbouring sequences in the third and fifth transmembrane segments also contribute to isoform-specific differences in anion selectivity. The conserved motif in the Cl⁻ channel P1 region may constitute a 'signature' sequence for an anion-selective ion pore by analogy with the homologous GYG sequence that is essential for selectivity in voltage-gated potassium ion (K⁺) channel pores^{6–8}.

All known ClC channels have a conserved motif located at the carboxy-terminal side of transmembrane segment D3 (designated as the P1 region) which in human ClC-1 (hClC-1) has the sequence GKEGPFVH (residues 230–237) (Fig. 1a). A naturally occurring substitution (G230E) in this segment that causes autosomal dominant myotonia congenita (Thomsen's Disease)⁹ greatly distorts permeation properties of the channel suggesting that this residue may lie within or near the ion-conduction pathway¹⁰. To investigate this region in more detail, we made amino-acid substitutions for each residue between Gly 230 and His 237, and evaluated channel function using the patch-clamp technique. Representative whole-cell current recordings and instantaneous current–voltage relationships obtained from each mutant are shown in Fig. 1b. Mutations in P1 affect open-channel rectification and block by extracellular iodide (Fig. 1b, open symbols). Furthermore, all mutations constructed in P1 confer altered anion selectivity on hClC-1 as deduced from permeability ratios (Table 1). Substitutions at positions 230,

231 and 233 have the greatest effect on the anion-selectivity sequence, whereas mutations at 232, 235 and 236 have smaller but significant effects on selectivity. In addition, mutations of Gly 230, Lys 231 and Gly 233 confer increased cation permeability (Table 1). Channel gating was also very sensitive to mutations of residues Gly 230 to Pro 234 as has been reported for pore mutations in voltage-gated K⁺ channels⁶.

To test whether residues in the P1 region project their side groups into an aqueous cavity, consistent with an ion-conduction pathway, we performed cysteine-scanning mutagenesis and examined the mutants for reactivity to substituted water-soluble methanethiol-sulphonates (MTS reagents)¹¹. Wild-type hClC-1 exhibits no functional change upon exposure to MTS reagents applied intra- or extracellularly under our experimental conditions (Fig. 2a). A cysteine placed at position 231 is accessible only to extracellularly applied MTS-ethylsulphonate (MTSES), whereas cysteines at positions 234 and 237 are reactive only to intracellularly applied MTSES (Fig. 2a). Application of MTS-ethyltrimethylammonium (MTSET) causes comparable reductions in current amplitudes (data not shown). The functional effects of MTS reagents on cysteine-sub-

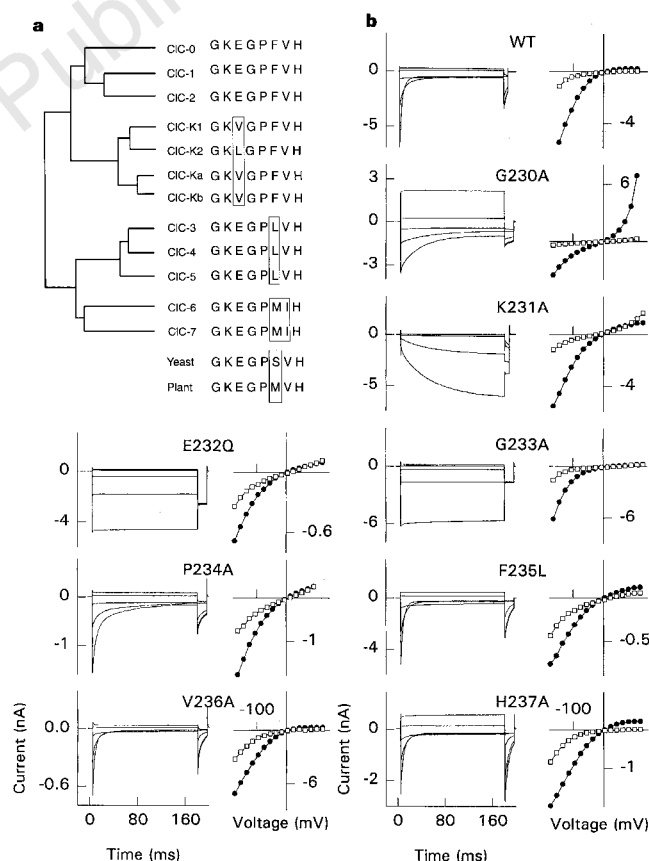


Figure 1 Evolutionary conservation of the GKEGPFVH motif in P1, and the effect of mutations. **a**, Alignment of P1 region in ClC channels. Least-conserved residues are boxed. **b**, Representative current recordings from WT and mutant hClC-1 channels. For each construct, voltage steps between –165 mV and +75 mV in 60-mV steps were applied from a holding potential of 0 mV. Each test pulse is followed by a fixed –105 mV pulse. Whole-cell recordings are shown for: WT, G230A, K231A, E232Q and G233A. Excised inside-out patches were used for: P234A, F235L, V236A, H237A. Instantaneous current–voltage relationships were measured after a 250 ms prepulse to +75 mV (–100 mV for K231A) in standard extracellular solution (filled circles) and for the same cell/patch after changing to (in mM): NaI (140), KCl (4), CaCl₂ (2), MgCl₂ (1), HEPES (5) (open squares). Whole-cell recordings are shown for: WT, G230A, K231A, G233A. Excised outside-out patches for: E232Q, P234A, F235L, V236A, H237A.

stituted mutants could be reversed with the disulphide-reducing reagent, dithiothreitol. These observations indicate that the segment between Lys231 and Pro234 spans an accessibility barrier between the external and internal solutions. Such a short span of amino-acid residues is not likely to cross the entire thickness of a lipid bilayer, and is more consistent with the lining of an aqueous pore near a constriction point presumed to exist in most ion-selective channels¹.

Determining the rate of modification by different MTS reagents can provide information about the charge selectivity of the environment surrounding the introduced cysteine¹². For both Cys231

and Cys237, calculated second-order rate constants for reduction in current amplitude are smaller than that observed for MTS modification of thiols in aqueous solution, suggesting that the access of the reagent to the reactive cysteine is rate-limiting. The charge selectivity of the access pathway for MTS reagents can therefore be assessed by comparing the reaction rate constants between anionic MTSES and cationic MTSET at a membrane potential of 0 mV (Fig. 2c). When the rate constant ratio ($k_{\text{MTSES}}:k_{\text{MTSET}}$) is normalized to the relative reactivity of the two reagents with thiols in aqueous solution (0.08)¹², we obtain values of 4.2 and 8.2 for K231C and H237C, respectively. We infer from these experimental results that

Table 1 Permeability ratios and ion selectivity sequences for hCIC-1 and mutants

	$P_{\text{I}}/P_{\text{Cl}}$	$P_{\text{SCN}}/P_{\text{Cl}}$	$P_{\text{NO}_3}/P_{\text{Cl}}$	$P_{\text{Br}}/P_{\text{Cl}}$	$P_{\text{Na}}/P_{\text{Cl}}$	Selectivity sequence
WT	0.34 ± 0.01	0.92 ± 0.04	0.53 ± 0.03	0.59 ± 0.06	0.10 ± 0.03	$\text{Cl} > \text{SCN} > \text{Br} > \text{NO}_3 > \text{I}$
G230A	2.22 ± 0.14	4.70 ± 0.03	1.89 ± 0.07	0.96 ± 0.02	$0.25 \pm 0.01^*$	$\text{SCN} > \text{I} > \text{NO}_3 > \text{Cl} = \text{Br}$
K231A	0.97 ± 0.02	3.79 ± 0.21	1.60 ± 0.13	0.92 ± 0.02	$0.36 \pm 0.04^*$	$\text{SCN} > \text{NO}_3 > \text{Cl} = \text{I} > \text{Br}$
K231E	1.31 ± 0.17	2.17 ± 0.30	1.72 ± 0.30	0.90 ± 0.04	$0.64 \pm 0.05^*$	$\text{SCN} > \text{NO}_3 > \text{I} > \text{Cl} > \text{Br}$
K231C	2.79 ± 0.10	5.48 ± 0.29	1.51 ± 0.04	1.05 ± 0.01	$0.36 \pm 0.04^*$	$\text{SCN} > \text{I} > \text{NO}_3 > \text{Br} > \text{Cl}$
E232Q	0.27 ± 0.06	1.34 ± 0.02	0.38 ± 0.02	0.60 ± 0.04	0.15 ± 0.08	$\text{SCN} > \text{Cl} > \text{Br} > \text{NO}_3 > \text{I}$
E232K	0.30 ± 0.05	1.42 ± 0.06	0.34 ± 0.04	0.61 ± 0.01	0.16 ± 0.07	$\text{SCN} > \text{Cl} > \text{Br} > \text{NO}_3 > \text{I}$
G233A	0.93 ± 0.01	7.8 ± 1.13	1.53 ± 0.20	0.66 ± 0.01	$0.25 \pm 0.04^*$	$\text{SCN} > \text{NO}_3 > \text{Cl} > \text{I} > \text{Br}$
P234A	0.34 ± 0.09	0.86 ± 0.02	0.52 ± 0.03	0.52 ± 0.02	ND	$\text{Cl} > \text{SCN} > \text{Br} = \text{NO}_3 > \text{I}$
F235L	0.13 ± 0.02	1.34 ± 0.07	0.25 ± 0.04	0.54 ± 0.01	ND	$\text{SCN} > \text{Cl} > \text{Br} > \text{NO}_3 > \text{I}$
V236A	0.27 ± 0.05	1.47 ± 0.02	0.21 ± 0.02	0.52 ± 0.01	ND	$\text{SCN} > \text{Cl} > \text{Br} > \text{I} > \text{NO}_3$
H237A	0.17 ± 0.02	1.65 ± 0.11	0.28 ± 0.05	0.46 ± 0.01	ND	$\text{SCN} > \text{Cl} > \text{Br} > \text{NO}_3 > \text{I}$
D265C	0.19 ± 0.03	1.28 ± 0.04	0.23 ± 0.01	0.54 ± 0.01	ND	$\text{SCN} > \text{Cl} > \text{Br} > \text{NO}_3 > \text{I}$
L267C	0.16 ± 0.01	1.20 ± 0.01	0.24 ± 0.02	0.44 ± 0.01	ND	$\text{SCN} > \text{Cl} > \text{Br} > \text{NO}_3 > \text{I}$
G285C	2.03 ± 0.10	5.75 ± 0.12	2.66 ± 0.11	1.65 ± 0.05	ND	$\text{SCN} > \text{NO}_3 > \text{I} > \text{Br} > \text{Cl}$
V286C	0.04 ± 0.01	1.33 ± 0.03	0.25 ± 0.02	0.41 ± 0.01	ND	$\text{SCN} > \text{Cl} > \text{Br} > \text{NO}_3 > \text{I}$
L287C	0.38 ± 0.06	1.49 ± 0.06	0.26 ± 0.01	0.49 ± 0.02	ND	$\text{SCN} > \text{Cl} > \text{Br} > \text{I} > \text{NO}_3$
F288C	0.22 ± 0.08	0.86 ± 0.01	0.26 ± 0.01	0.55 ± 0.01	ND	$\text{Cl} > \text{SCN} > \text{Br} > \text{NO}_3 > \text{I}$
S289C	0.12 ± 0.02	1.23 ± 0.03	0.20 ± 0.02	0.50 ± 0.03	ND	$\text{SCN} > \text{Cl} > \text{Br} > \text{NO}_3 > \text{I}$
I290C	0.10 ± 0.02	1.55 ± 0.01	0.19 ± 0.01	0.53 ± 0.03	ND	$\text{SCN} > \text{Cl} > \text{Br} > \text{NO}_3 > \text{I}$

* $P < 0.05$ (Student's *t*-test) when compared to wild type (WT). Data shown are mean $\pm$ s.e.m. ($n = 4$). WT, Wild-type hCIC-1. ND, not determined.

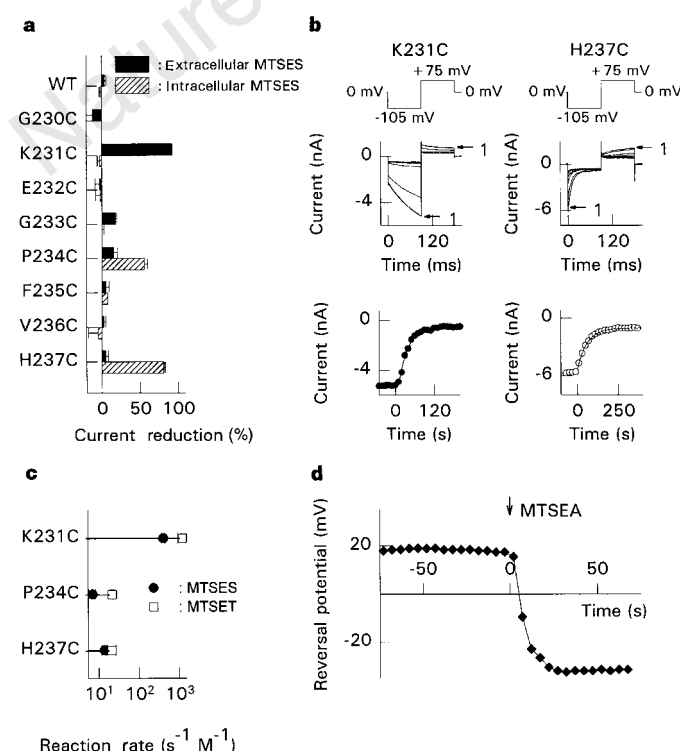


Figure 2 Substituted cysteine accessibility experiments of P1 region. **a**, Relative current reduction for each substituted cysteine mutant following application of either intracellular (hatched bars) or extracellular (solid bars) MTSES. **b**, Current recordings during modification of K231C (0.1 mM MTSES) and H237C (1 mM MTSES) hCIC-1 mutants. Whole-cell recordings of K231C, and excised inside-out recordings of H237C are shown. The pulse protocol consists of a voltage step to -105 mV followed by $+75$ mV step. A holding potential of 0 mV was applied between trials. The lower graphs show the change of the instantaneous current amplitude measured at -105 mV with time. The solid line represents a fit with a monoexponential time dependence. **c**, Reaction rates for modification of Cys231, Cys234 and Cys237 by MTSES or MTSET. Second-order rate constants measured at 0 mV for MTS modification of K231C were: $k_{\text{omv}} = 405.6 \pm 31.4 \text{ s}^{-1} \text{ M}^{-1}$ ($n = 4$) for extracellularly applied MTSES, and $k_{\text{omv}} = 1197.9 \pm 272 \text{ s}^{-1} \text{ M}^{-1}$ ($n = 4$) for extracellularly applied MTSET. Rate constants for P234C: $k_{\text{omv}} = 7.1 \pm 1.4 \text{ s}^{-1} \text{ M}^{-1}$ ($n = 4$) for intracellularly applied MTSES, and $k_{\text{omv}} = 21.6 \pm 1.0 \text{ s}^{-1} \text{ M}^{-1}$ ($n = 4$) for intracellularly applied MTSET. Rate constants for H237C: $k_{\text{omv}} = 13.9 \pm 2.1 \text{ s}^{-1} \text{ M}^{-1}$ ($n = 4$) for intracellularly applied MTSES, and $k_{\text{omv}} = 20.9 \pm 5.9 \text{ s}^{-1} \text{ M}^{-1}$ ($n = 4$) for intracellularly applied MTSET. **d**, Changes of the current reversal potential with time during MTSEA modification of K231C. Cells were dialysed with an internal solution consisting of (in mM): NaI (132), EGTA (5), HEPES (10) and bathed in standard extracellular solution. The current reversal potential was monitored by using the current-clamp mode at a 1 Hz frequency. $5 \mu\text{M}$ MTSEA was applied at time 0 .

the access pathway in the cysteine-substituted mutants for MTS reagents is intrinsically anion-selective.

Replacement of Lys 231 with cysteine causes reversal of the $\text{Cl}^- > \text{I}^-$ permeability sequence characteristic of WT-hClC-1 (Table 1). To examine the importance of positive charge at this position for anion selectivity, we tested whether adding a small cationic adduct to Cys 231 would restore selectivity to the wild-type pattern. Figure 2d shows the effect of 5 μM extracellular MTSEA on the reversal potential measured in cells expressing K231C under biionic conditions with Γ^- as the principal intracellular anion and Cl^- as the principal extracellular anion. The current reversal potential changes with a monoexponential time course after the addition of MTSEA from $16.6 \pm 1.8 \text{ mV}$ ($n = 7$) to $-18.0 \pm 2.9 \text{ mV}$, corresponding to $P_{\text{I}}/P_{\text{Cl}}$ permeability ratios of 2.17 ± 0.15 (before) and 0.60 ± 0.07 (after modification), respectively. The $P_{\text{I}}/P_{\text{Cl}}$ value observed for MTSEA-modified K231C indicates a $\text{Cl}^- > \text{I}^-$ selectivity and is very close to the value observed for the wild-type channel ($P_{\text{I}}/P_{\text{Cl}} = 0.34$). No MTSEA reactivity was observed with the K231A mutant which has similar gating and anion-selectivity properties.

Lysine 231 is also critical for the low-cation selectivity of hClC-1. Charge inversion at position 231 (K231E) causes an increased cation permeability as compared with wild-type channels, and has a greater effect than the charge neutralizing mutation, K231A (Table 1). By contrast, side group charge at position 232 does not appear to influence ion selectivity as one might expect for a residue near or within the conduction pathway (Table 1). This observation along with the apparent inaccessibility of Glu 232 to MTSES (Fig. 2a) suggests that the negatively charged side group of Glu 232 projects away from the ionic pore.

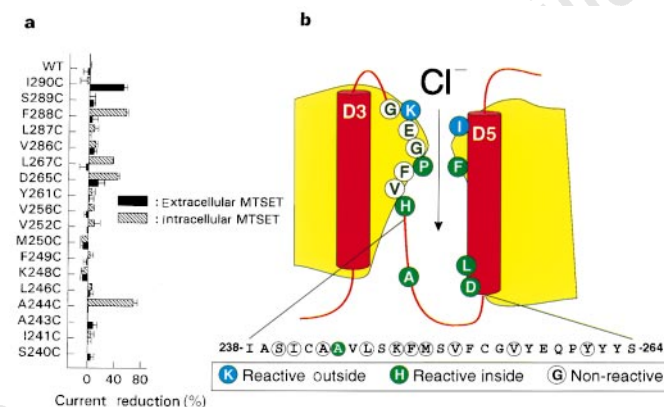


Figure 3 Substituted cysteine accessibility experiments of D5 and revised transmembrane topology for the ClC channel pore. **a**, Relative current reduction for each substituted cysteine mutant 3 min after application of 1 mM MTSET either to the cytoplasmic side of excised inside-out (hatched bars) or to the extracellular side of excised outside-out patches (solid bars). Cysteine mutants I290C, S289C, F288C, L287C, V286C, L267C and D265C are in D5. The remaining cysteine mutants are located in the P1-D5 linker. **b**, Model of D3-D5 region indicating MTS reagent accessibility of residues. Cysteine substitutions were made for each amino-acid position shown enclosed by a circle. MTS reactivity is indicated by the colour shading as described in the inset legend.

The P1 region alone cannot account for isoform-specific patterns of ion selectivity. Rat ClC-3 exhibits permeation properties that are distinct from hClC-1¹³, but a single amino-acid difference in the P1 region between the two isoforms does not confer the ClC-3 pattern of anion selectivity upon hClC-1 (F235L, Table 1). We used a chimera strategy to test whether regions adjacent to P1 are important for isoform-specific ion selectivity. A chimera in which sequences between the D2-D3 loop and the C-terminal end of D5 of hClC-1 were replaced by the corresponding region of ClC-3 (designated D3-D5) exhibits the same ion selectivity sequence as ClC-3 (Table 2). This suggests that residues conferring isoform-specific pore properties are present in the D3-D5 region. We constructed three additional chimaeras in which segments of this region of hClC-1 were replaced by the corresponding structures from ClC-3 (Table 2). The D3N chimera exhibits the least altered anion-selectivity sequence and we conclude that the D2-D3 loop and N-terminal portion of D3 are not critical determinants of isoform-specific ion selectivity. By contrast, both D3C-P1 and D5 chimaeras exhibit changes in ion selectivity that resemble ClC-3. Therefore, sequences within the C-terminal portion of D3 and D5 regions are important determinants of isoform-specific ion selectivity. The ability of these regions to confer isoform-specific patterns of ion selectivity strongly suggests that these structures are involved in the formation of the channel pore.

The C-terminal portion of D5 contains a highly conserved sequence motif (GVLFISI in hClC-1) and we used cysteine-scanning mutagenesis to support the hypothesis that this region also contributes to the ion-conduction pathway. Cysteine substitution of Gly 285 has a large effect on the anion-selectivity sequence whereas substitutions of other residues within this motif have slight effects (Table 1). Cysteine at position 290 (I290C) renders the channel susceptible to functional modification by 1 mM extracellular but not intracellular MTSET, whereas current amplitude in cells expressing the F288C mutation is reduced upon exposure to 1 mM intracellular, but not extracellular, MTSET (Fig. 3a). Thus the GVLFISI motif is important for anion selectivity and spans an accessibility barrier between the external and internal solutions, consistent with a pore-forming segment.

The internal accessibility of Cys 288 and the external accessibility of Cys 290 contradicts previously suggested topological models of ClC channels^{14,15}. To define further the transmembrane orientation of P1-D5, we tested accessibility of cysteine residues placed within the N-terminal portion of D5, and within the P1-D5 linker region (formerly known as 'D4'). Cysteine substitution at positions 265, 267 (N-terminal D5) and 244 (P1-D5 linker) renders the channel susceptible to block by intracellular, but not extracellular MTSET (Fig. 3a). Ten additional cysteine substitutions within P1-D5 exhibited no functional change when exposed to either intra- or extracellular MTSET. These results are entirely consistent with a new model for the D3-D5 region illustrated in Fig. 3b.

In summary, the highly conserved P1 region is a critical structural determinant of ion selectivity, and in conjunction with portions of D3 and D5, forms a substantial part of the ClC ionic pore. Two cationic amino-acid residues (Lys 231, His 237) project into the ion-conduction pathway amidst neighbouring hydrophobic structures

Table 2 Permeability ratios and ion-selectivity sequences for hClC-1/ClC-3 chimaeras

	D2	D3	P1	D5	$P_{\text{SCN}}/P_{\text{Cl}}$	$P_{\text{I}}/P_{\text{Cl}}$	$P_{\text{NO}_3}/P_{\text{Cl}}$	$P_{\text{Br}}/P_{\text{Cl}}$	Selectivity sequence
hClC-1					0.92 ± 0.04	0.34 ± 0.01	0.53 ± 0.03	0.59 ± 0.06	$\text{Cl}^- > \text{SCN}^- > \text{Br}^- > \text{NO}_3^- > \text{I}^-$
D3-D5					4.75 ± 0.19	2.43 ± 0.17	2.71 ± 0.16	1.64 ± 0.07	$\text{SCN}^- > \text{NO}_3^- > \text{I}^- > \text{Br}^- > \text{Cl}^-$
D3N					1.37 ± 0.02	0.10 ± 0.02	0.29 ± 0.01	0.56 ± 0.08	$\text{SCN}^- > \text{Cl}^- > \text{Br}^- > \text{NO}_3^- > \text{I}^-$
D3C-P1					8.42 ± 1.39	2.90 ± 0.17	1.50 ± 0.05	1.20 ± 0.01	$\text{SCN}^- > \text{I}^- > \text{NO}_3^- > \text{Br}^- > \text{Cl}^-$
D5					18.70 ± 1.60	2.97 ± 0.11	5.05 ± 0.54	2.05 ± 0.10	$\text{SCN}^- > \text{NO}_3^- > \text{I}^- > \text{Br}^- > \text{Cl}^-$
ClC-3					4.46 ± 1.03	2.13 ± 0.11	2.33 ± 0.13	1.44 ± 0.16	$\text{SCN}^- > \text{NO}_3^- > \text{I}^- > \text{Br}^- > \text{Cl}^-$

Chimaera D3-D5: replacement of hClC-1 residues 194-290 by ClC-3 186-279; chimaera D3N: replacement of hClC-1 residues 194-210 by ClC-3 186-201; chimaera D3C-P1: replacement of hClC-1 residues 216-246 with ClC-3 208-238; chimaera D5: replacement of hClC-1 residues 250-290 with ClC-3 240-279.

and may serve as anion-binding sites within the channel pore¹⁶. This combination of cationic and hydrophobic groups may contribute to the experimentally observed greater binding affinities of large and polyatomic anions for the hClC-1 pore¹⁷. □

Methods

Mutagenesis. Site-specific mutants and chimaeric channels were constructed using recombinant polymerase chain reaction (PCR) mutagenesis¹⁸. All constructs were assembled in the expression construct pRc/CMV-hClC-1 (ref. 19) and regions modified by PCR were sequenced completely to exclude polymerase errors. At least two independent recombinants were examined for each mutant or chimaera. Transient transfection of tsA201 cells was performed as previously described¹⁰.

Electrophysiology. Standard whole-cell, inside-out or outside-out patch clamp recordings were performed²⁰ using experimental conditions as previously described²¹. The standard extracellular solution contained (in mM): NaCl (140), KCl (4), CaCl₂ (2), MgCl₂ (1), HEPES (5), pH 7.4, and the standard intracellular solution contained (in mM): NaCl (130), MgCl₂ (2), EGTA (5), HEPES (10), pH 7.4. Reversal potentials were measured in the current-clamp mode using the whole-cell configuration. For determining the relative anion selectivity, cells were internally perfused with standard solution and bathed in an extracellular medium containing (in mM): NaX (140), KCl (4), CaCl₂ (2), MgCl₂ (1), HEPES (5), pH 7.4, where X represents different anions. Solution changes were accomplished by moving the cell/patch into the stream of a silanized-treated macropipette filled with the test solution. For determining relative cation selectivity, cells held in the whole-cell configuration were moved quickly into a solution stream containing (in mM): NaCl (300), HEPES (5) pH 7.4. The reversal potential was obtained immediately after placing the cell in the solution stream. Using the different reversal potentials, P_X/P_{Cl} (X denotes SCN⁻, I⁻, NO₃⁻, Br⁻ or Na⁺) were then calculated using the Goldman-Hodgkin-Katz equation¹. Data were analysed by a combination of pClamp (Axon Instruments, Foster City, CA) and SigmaPlot (Jandel Scientific, San Rafael, CA) programs. All data are shown as means ± s.e.m. for at least three cells.

Modification with MTS reagents. 2-Aminoethyl-methanethiosulphonate (MTSEA), 2-(trimethylammonium)ethyl-methanethiosulphonate (MTSET), and 2-sulphonatoethyl-methanethiosulphonate (MTSES) were obtained from Toronto Research Chemicals (North York, Ontario). Stock solutions (0.1 M) were prepared in distilled water, stored at -20 °C, and diluted into the bath solution immediately before use. Cells/patches were stimulated as shown in Fig. 2b. Typically, 20 cycles were used to assess control values (I_{control}), the MTS-reagents were applied to cells/patches by moving the cell/patch into the stream of a silane-treated macropipette filled with MTS-containing solution with concentrations of 1 mM MTSET and 10 mM MTSES. Extracellular reagent concentrations were reduced for K231C experiments (50 μM MTSET, 100 μM MTSES). After 3 min, cells/patches were moved out of the stream to visualize reversible effects and to measure the current amplitude after irreversible modification (I_{modified}). The relative current reduction was measured as $1 - I_{\text{modified}}/I_{\text{control}}$. For all cysteine mutants except G230C, K231C, G233C and G285C, excised inside-out and outside-out patches were used. The time course of modification was fitted with a single exponential giving the time constant of modification. The pseudo-first-order rate constant was calculated as the inverse of the modification time constant. Dividing by the concentration of the MTS reagents provided the second-order rate constants.

Received 11 June; accepted 9 September 1997.

- Hille, B. *Ionic Channels of Excitable Membranes* (Sinauer, Sunderland, MA, 1992).
- Jentsch, T. J. Molecular biology of voltage-gated chloride channels. *Curr. Top. Membr.* **42**, 35–57 (1994).
- Steinmeyer, K. *et al.* Inactivation of muscle chloride channel by transposon insertion in myotonic mice. *Nature* **354**, 304–308 (1991).
- Koch, M. C. *et al.* The skeletal muscle chloride channel in dominant and recessive human myotonia. *Science* **257**, 797–800 (1992).
- Lloyd, S. E. *et al.* A common molecular basis for three inherited kidney stone diseases. *Nature* **379**, 445–449 (1996).
- Heginbotham, L., Lu, Z., Abramson, T. & MacKinnon, R. Mutations in the K⁺ channel signature sequence. *Biophys. J.* **66**, 1061–1067 (1994).
- Ranganathan, R., Lewis, J. H. & MacKinnon, R. Spatial localization of the K⁺ channel selectivity filter by mutant cycle-based structure analysis. *Neuron* **16**, 131–139 (1996).
- Lu, Q. & Miller, C. Silver as a probe of pore-forming residues in a potassium channel. *Science* **268**, 304–307 (1995).
- George, A. L., Crackower, M. A., Abdalla, J. A., Hudson, A. J. & Ebers, G. C. Molecular basis of Thomsen's disease (autosomal dominant myotonia congenita). *Nature Genet.* **3**, 305–310 (1993).

- Fahlke, Ch., Beck, C. L. & George, A. L. Jr A mutation in autosomal dominant myotonia congenita affects pore properties of the muscle chloride channel. *Proc. Natl. Acad. Sci. USA* **94**, 2729–2734 (1997).
- Akabas, M. H., Stauffer, D. A., Xu, M. & Karlin, A. Acetylcholine receptor channel structure probed in cysteine-substitution mutants. *Science* **258**, 307–310 (1992).
- Cheung, M. & Akabas, M. H. Locating the anion-selectivity filter of the cystic fibrosis transmembrane conductance regulator (CFTR) chloride channel. *J. Gen. Physiol.* **109**, 289–299 (1997).
- Kawasaki, M. *et al.* Cloning and expression of a protein kinase C-regulated chloride channel abundantly expressed in rat brain neuronal cells. *Neuron* **12**, 597–604 (1994).
- Middleton, R. E., Pheasant, D. J. & Miller, C. Purification, reconstitution, and subunit composition of a voltage-gated chloride channel from *Torpedo* electroplax. *Biochemistry* **33**, 13189–13198 (1994).
- Schmidt-Rose, T. & Jentsch, T. J. Transmembrane topology of a CLC chloride channel. *Proc. Natl. Acad. Sci. USA* **94**, 7633–7638 (1997).
- Dani, J. A., Sanchez, J. A. & Hille, B. Na channel gating and Ca electrode response. *J. Gen. Physiol.* **81**, 255–281 (1983).
- Fahlke, Ch., Dürre, C. & George, A. L. Jr Mechanism of ion permeation in skeletal muscle chloride channels. *J. Gen. Physiol.* **110**, 551–564.
- Higuchi, R. in *PCR Technology* (ed. Erlich, H. A.) 61–70 (Stockton Press, New York, 1989).
- Fahlke, Ch., Rüdel, R., Mitrovic, N., Zhou, M. & George, A. L. Jr An aspartic acid residue important for voltage-dependent gating of human muscle chloride channels. *Neuron* **15**, 463–472 (1995).
- Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. *Pflügers Arch.* **391**, 85–100 (1981).
- Fahlke, Ch., Rosenbohm, A., Mitrovic, N., George, A. L. Jr & Rüdel, R. Mechanism of voltage-dependent gating in skeletal muscle chloride channels. *Biophys. J.* **71**, 695–706 (1996).

Acknowledgements. We thank S. Uchida for providing rat ClC-3, M. Akabas for helpful discussions, L. Limbird and D. Snyder for critical reviews of the manuscript. We also thank R. Desai and C. Ryan for technical assistance. This work was supported by the Muscular Dystrophy Association (A.L.G., Ch.F.), the Lucille P. Markey Charitable Trust (A.L.G.) and the German Research Foundation (Ch.F.). C.L.B. is a recipient of the Luis and Emma Benzak Neuromuscular Disease Research Fellowship from the Muscular Dystrophy Association. H.T.Y. is a recipient of a National Kidney Foundation Research Fellowship.

Correspondence and requests for materials should be addressed to A.L.G. (e-mail: al.george@mcmill.vanderbilt.edu).

Ammonia mediates communication between yeast colonies

Zdena Palková*, Blanka Janderová*, Jiří Gabriel†, Blanka Zikánová*, Martin Pospíšek* & Jitka Forstová*

* Department of Genetics and Microbiology, Charles University, Viničná 5, 12844 Prague 2, Czech Republic

† Institute of Microbiology, Academy of Sciences of the Czech Republic, Videňská 1083, 14220 Prague 4, Czech Republic

Under certain growth conditions unicellular organisms behave as highly organized multicellular structures. For example, the fruiting bodies of myxobacteria¹ and of the slime mould *Dictyostelium discoideum*² form structures composed of non-dividing motile cells. Although non-motile, yeasts can create organized structures, colonies in which cells communicate and act in a coordinated fashion. Colony morphologies are characteristic for different species and strains. Here we describe that, in addition to short-range intracolony cell–cell communication, yeasts exhibit long-distance signals between neighbouring colonies. The volatile alkaline compound ammonia, transmitted by yeast colonies in pulses, has been identified as a substance mediating the inter-colony signal. The first alkaline pulse produced by neighbouring colonies is non-directed and is followed by acidification of the medium. The second pulse seems to be enhanced and is oriented towards the neighbour colony. Ammonia signalling results in growth inhibition of the facing parts of both colonies. This phenomenon is observed in different yeast genera. The presence of amino acids in the medium is required for ammonia production. Colonies derived from the yeast *Saccharomyces cerevisiae* *shr3* mutant, defective in localization of amino-acid permeases³, do not produce detectable amounts of ammonia and do not exhibit asymmetric growth inhibition.

We observed that when yeast colonies grow on complex agar medium (GM agar) with high concentrations of Ca²⁺ ions, turbid paths directed from one colony to another were formed (Fig. 1a).

These oriented turbid zones appeared after several days of growth (depending on the yeast strain) in the surface layer of the agar. Similar zones were observed among the 'giant' colonies (Fig. 1b–g) arising from more than one cell (from a cell suspension spotted onto agar)⁴. Like colonies originating from one cell, 'giant' colonies have a highly organized morphology specific for a particular yeast strain. After the turbid path formation, the growth of neighbouring colonies became inhibited in their facing borders (Fig. 1g).

The formation of turbid paths between colonies appears to be a common phenomenon. Colonies of nine analysed genera (*Candida*, *Cryptococcus*, *Endomyces*, *Hansenula*, *Kluyveromyces*, *Rhodotoridium*, *Rhodotorula*, *Saccharomyces*, *Schwanniomyces*) created oriented zones on the same medium, although the timing in their appearance differed, probably as a consequence of different growth rates (see Fig. 1). Turbid paths were also observed between different genus colonies (for example, between *Schwanniomyces occidentalis* and *Candida mogii*; data not shown).

Ca^{2+} is necessary for turbid path formation and cannot be replaced by other bivalent (Mg^{2+}) or univalent cations (Na^+ , K^+). The turbid path can be solubilized by EDTA. The absence of turbid paths between colonies growing on GM agar lacking Ca^{2+} does not prevent the asymmetric inhibition of their growth. Thus, Ca^{2+} ion-mediated precipitation is an artefact and not essential for growth inhibition. The use of different carbon sources (glycerol, glucose, acetate, pyruvate, lactate) in Ca^{2+} -rich GM agar did result in differences in the timing of the appearance of the turbid path, probably due to their effects on the growth rate. No turbid zones developed on minimal agar (MA) medium, even when enriched

with Ca^{2+} . The zones appeared when MA medium was supplemented either with yeast extract or casamino acids, indicating that an amino acid(s) participates in their formation.

We set out to answer two questions: How do yeast colonies recognize their neighbours? What signals are transmitted and received? Using various barriers between colonies we concluded that the primary signal is a small, volatile compound. Two colonies growing on agar discs separated by an air-layer (Fig. 2) were still able to create turbid paths, pointing towards each other.

pH changes around a single growing giant colony, or two giant colonies, were followed by a dye indicator (bromocresol purple) added to the GM agar (Fig. 3). When two giant colonies were followed, two pulses of alkali (appearance of violet colour around colonies), separated by a peak of acidification (yellow colour around colonies), were observed during colony development. The second alkaline peak was the more intense of the two and appeared initially between colonies. The timing of pH changes depended on the strain used; the start of the second alkali pulse always preceded turbid zone formation. Decreased Ca^{2+} concentrations in the medium did not affect pH changes, but prevented zone formation (data not shown). Solitary colonies exhibited significantly lower intensities of the second alkaline peak and a substantial delay in its appearance (Fig. 3, compare 94 and 158 h). The observation of transient alkali gradients around growing colonies was surprising because the medium is reportedly continuously acidified during the submersion cultivation of yeasts^{5,6}.

We absorbed the volatile signalling compound produced by yeast colonies into small vessels containing an acidic solution (either 10% citric acid or 2% HCl) placed on the lid of an inverted dish below the growing giant colonies (Fig. 4Aa). Analysis of the solution by Nessler's reagent, and by high-performance liquid chromatography (HPLC) identified the absorbed compound as ammonia. No ammonia (NH_3) was detected in the absence of yeast colonies. The time course of NH_3 production by two giant *C. mogii* colonies (Fig. 4Ab), measured for 7 days, exhibited two peaks, as was also observed previously when measuring pH changes in medium

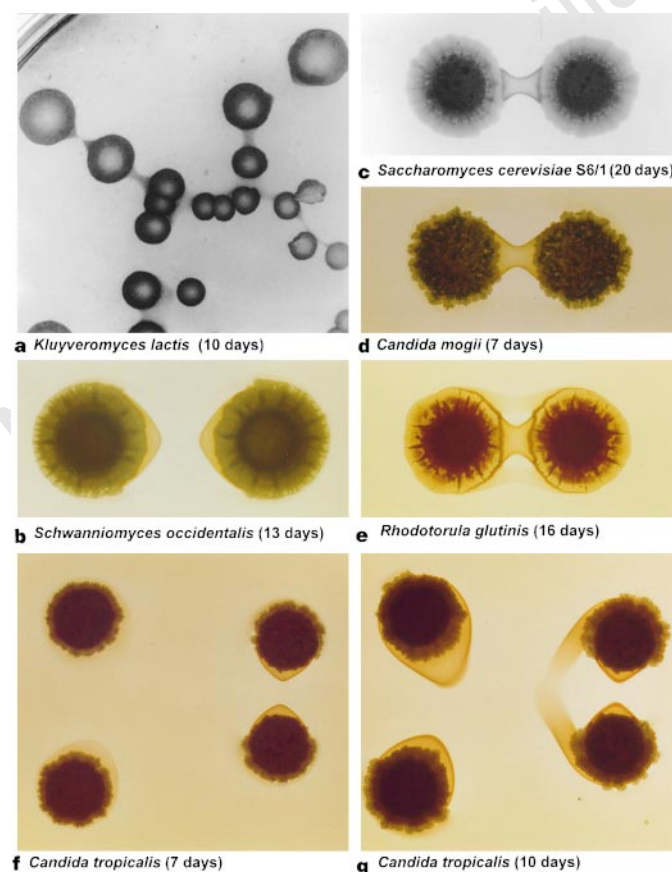


Figure 1 Oriented turbid zones between yeast colonies (a) and 'giant colonies' (b–g) on GM agar. As shown with *C. tropicalis*, the first turbid zone appeared between the nearest colonies (see f). Later, secondary zones developed between more distant colonies (g), emanating from the still growing border of the colony. No growing border can be seen in the area where the primary zone originated.

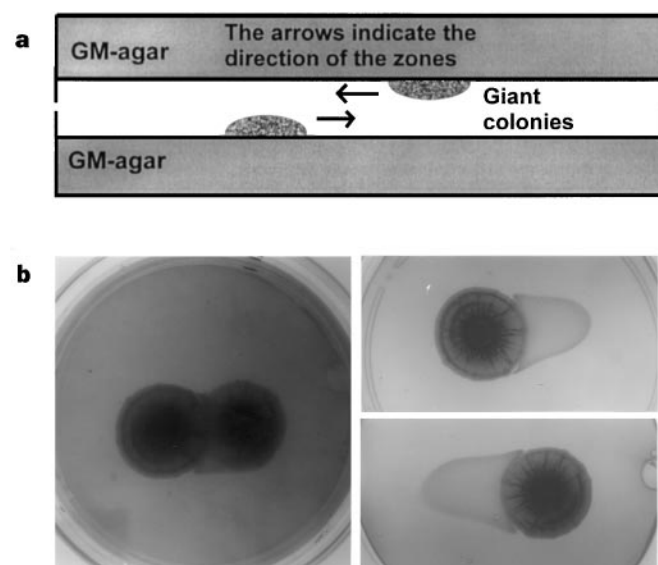


Figure 2 A volatile compound mediates intercolony signalling. a, Schematic diagram of the experiment where two yeast colonies, physically separated on agar layers, on the top and bottom of a Petri dish, are allowed to communicate only through an air layer. b, Experimental data, with two 'giant' colonies of *S. cerevisiae* GRF18, growing for 20 days on GM agar (on the bottom and lid of a plate), separated by an air layer, can be seen to form turbid zones in the direction pointing towards each other. The left panel shows a picture of the closed plate, and right panels show the separate lid and bottom of this plate.

surrounding yeast colonies over time (Fig. 3). The second increase of NH_3 production was more rapid than that obtained with a single growing colony (Fig. 4Ab and c). Efficient absorption of NH_3 , accomplished by exchanging the acidic trap (Fig. 4Aa) during colony growth, reduced the second peak of NH_3 production and prevented turbid path formation between colonies. To confirm that NH_3 diffusion (and/or the resulting alkali gradient) is responsible for induction of turbid path formation between neighbouring colonies, an artificial source of NH_3 was substituted for a partner colony. This experiment (Fig. 4B) showed that NH_3 mimicked the presence of the neighbouring colony, inducing turbid path formation from the solitary colony. It also inhibited growth of this colony on the side towards the NH_3 source (Fig. 4Bb and c).

To find mutants with defective intercolony NH_3 signalling, we analysed various *S. cerevisiae* mutants affected in cell-cycle and cell-

signalling pathways. The following mutations were tested: *ras2* (ref. 7); *ras2*^{Val19} (ref. 7); *dpr1* (ref. 8); *sst2* (ref. 9); *bar1* (ref. 9); *rml1* (ref. 10) and *shr3* (ref. 3), two of which showed impaired signalling. The strain with the *dpr1* mutation, defective in processing of Ras proteins and other substrates⁸, exhibited a substantial delay in turbid path formation compared with the parental strain. Strain 3639, carrying the *shr3* mutation, completely failed to create turbid zones (Fig. 5b). Also, the juxtaposition of its colony and a *SHR3* wild-type colony (strain 3640) did not result in formation of a turbid path between them (data not shown). This defect was recessive, manifested only by the 3639 strain, homozygous for the mutated *shr3* allele; the isogenic heterozygous 3638 strain, similar to the 3640 parental strain, created turbid zones without any delay. Colonies of mutant 3639 produced a minimum of NH_3 (if any) at the time of the first pulse observed in wild-type strain 3640 (1–6

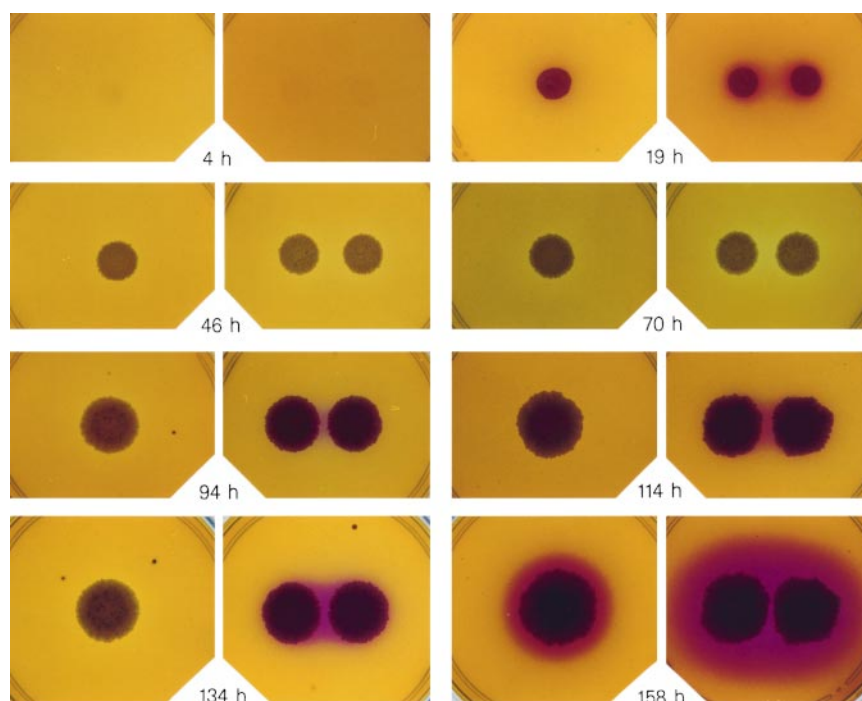


Figure 3 pH changes during *C. mogii* colony development over time. Data shown are for single or two partner colonies. Bromocresol purple in GM agar changes colour from violet (above pH 6.8) to yellow (below pH 5.2).

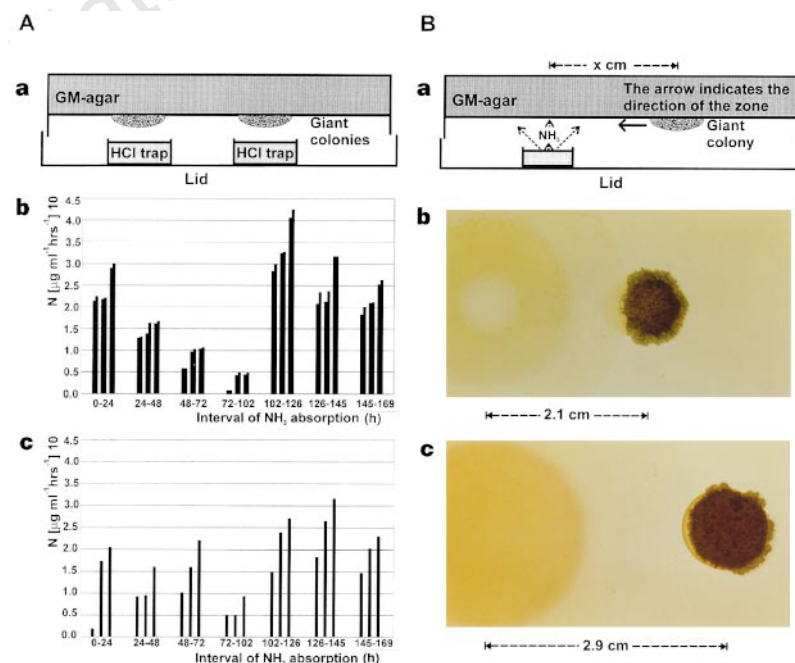


Figure 4 Ammonia generated in pulses by yeast colonies (**A**), and inhibiting their growth (**B**). **A**, NH_3 production by *C. mogii* giant colonies. Design of the experiment: plates with GM-agar (seven in each of three parallel experiments) were inoculated to allow development of two giant colonies each. Acidic traps were placed into each plate underneath respective colonies (as indicated in **a**) for a given time interval during which plates were incubated undisturbed. The amounts of NH_3 absorbed in individual traps were determined (**b**). The values obtained from traps placed in one plate are given separately. **c**, Concentrations of NH_3 produced by solitary colonies in the same time intervals. **B**, Solitary *C. mogii* colonies growing in the proximity of an ammonia source. **a**, Experimental design; **b**, **c**, growth of two individual colonies, placed at indicated distances from the centre of the NH_3 source.

days); no production was detected during the time of the second pulse (15–20 days) (Fig. 5a). Accordingly, the pH dye indicator did not detect any increase of pH around the mutant colonies. Colony morphology of strain 3639 differed dramatically from that of parental strain 3640, displaying an amorphous rather than defined pattern (Fig. 5c). The *SHR3* gene encodes a protein, located in the endoplasmic reticulum, that is responsible for the correct localization of several yeast amino-acid permeases⁵.

The observation that addition of casamino acids to MA medium restored turbid zone formation, together with the failure of the *shr3* mutant (defective in amino-acid uptake) to produce ammonia, suggested that amino acids might be both required for turbid path formation, and essential for ammonia signalling. Therefore, the behaviour of yeast colonies (*C. mogii*) on MA medium and MA medium supplemented with casamino acids (MACA) was compared. On the MA medium (without casamino acids) no turbid path formation between neighbouring colonies was seen and no induction of the second ammonia pulse production by partner colonies was observed (not shown). Colonies exhibited altered (less organized) morphologies and, because of the absence of the mutual growth inhibition, they overgrew each other (Fig. 6a). The absence of asymmetric growth inhibition on poor MA medium suggests that the growth inhibition is an active process rather than merely a consequence of nutrient depletion. Colonies grown on MACA medium showed substantially higher production of ammonia and stronger asymmetry of growth as compared with colonies grown on the complex GM medium (Fig. 6b). On the contrary, addition of increased amounts of NH_4Cl to both minimal (MA) and complex (GM) media, did not enhance the ammonia production and asymmetric growth inhibition by partner colonies (not shown). From these observations we conclude that the production of ammonia by yeast cells is probably connected with uptake and/or degradation pathway(s) of amino acids.

Two permeases (*MEP1* and *MEP2*) that transport ammonia across the plasma membrane into yeast cells have been identified¹¹. No data are available on whether *MEP* permeases are affected by *shr3* mutations or on the export of ammonia from yeast cells.

Ammonia has been identified as a signaling molecule involved in later stages of development of *Dictyostelium discoideum*, where it maintains *Dictyostelium* in the slug stage, preventing premature culmination by repressing a cAMP accumulation¹².

Our results indicate that NH_3 can act as a long-distance regulator of growing yeast colonies. Arguments supporting the involvement of ammonia as a long-distance intercellular signal in yeast growth also come from studies of pseudohypha growth of *Saccharomyces cerevisiae*. The switch of the yeast to filamentous pseudohyphal form of growth can occur in response to nitrogen starvation¹³. Pseudohyphae can respond to a nitrogen source concentration gradient by rapidly switching from the pseudohyphal to the yeast growth form¹⁴. Inhibition of pseudohyphae formation was also observed at higher colony densities¹⁴. We explain this inhibition by enhanced NH_3 production in regions of dense colony populations.

We suggest that yeast colonies produce ammonia, creating a temporary alkaline gradient, which is perceived as a signal by a neighbouring colony and, in response, NH_3 production is amplified in a retrodirection. Consequently, the growth in the direction of amplified signals is inhibited. In this way, colonies orient their growth towards areas that minimize the competition for nutrients.

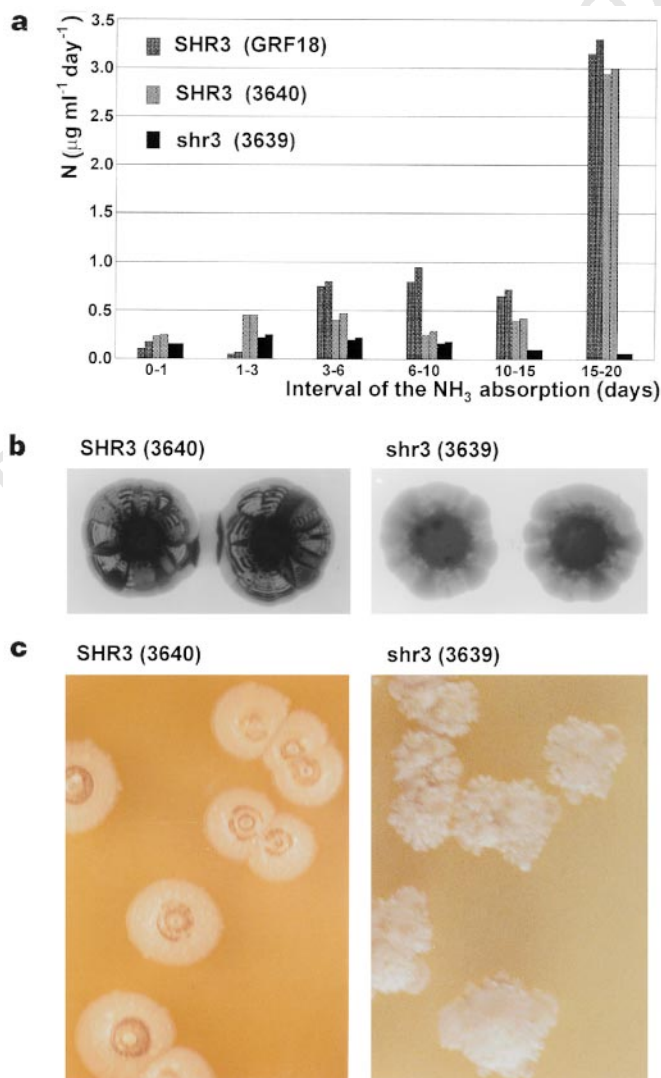


Figure 5 Ammonia production, turbid zone formation and colony morphology of the *S. cerevisiae* *shr3* mutant and *SHR3* strains. **a**, Time course measurement of NH_3 production; **b**, zone formation between giant colonies; **c**, morphology of colonies arising from a single cell. Colonies were grown on GM-agar.

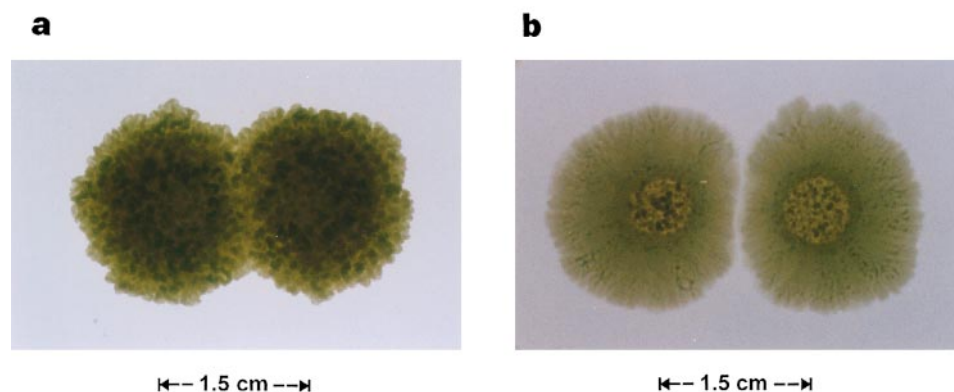


Figure 6 Giant colonies of *C. mogii* growing on MA medium (**a**) or on MACA medium (**b**). Partner colonies were inoculated at 1.5 cm distances.

Methods

Media and yeast strains. GM-agar (1% yeast extract, 3% glycerol, 2% agar, 30 mM CaCl₂); GM-BKP agar (GM-agar, 0.01% bromocresol purple); minimal agar MA (0.1% KH₂PO₄, 0.5% (NH₄)₂SO₄, 0.05% MgSO₄, 3% glycerol, 0.1% Wickerham's vitamin solution¹⁵, 2% agar); minimal agar with casamino acids MACA (MA, 0.5% casamino acids). Cultivation temperature was 28 °C.

Strains of *S. cerevisiae* 3638 (CGX15 a/α *shr3-102/SHR3 ura3-52/URA3*), 3639 (CGX19 a/α *shr3-102/shr3-102 ura3-52/ura3-52*), 3640 (CGX31 a/α *ura3-52/ura3-52*); 1505 (a *dpr1-1 his3 leu2 ura3 trp1 ade can1 RAS2^{Val19}*); 2700 (RC757 α *sst2 rme1 his6 met1 can1 cyh2*); 2112 (a *bar1::HISG ura3 leu2-3 112 trp1-1 his3 ade2-1 met*) were kindly provided by K. Nasmyth (Institute of Molecular Pathology, Vienna). Strains of *S. cerevisiae* KT131 (α *ura3 leu2 ras2::LEU2 trp1 his3 lys1 lys2*), KT301 (α *ura3 leu2 RAS2^{Val19} trp1 his3 lys1 lys2*) were kindly provided by A. Pichová (Institute of Microbiology, Prague). Strain GRF18 (α *his3 leu2*) and other wild-type yeast strains were from the Collection of Yeast Cultures of the Department of Genetics and Microbiology (DMUP), Charles University, Prague.

Photography. Colonies were photographed either with illuminating light coming through a plate from the bottom (Figs 1, 2, 3, 4, 5b, 6) or they were illuminated from above (Fig. 5c).

Ammonia detection and concentration measurement. Volatile compounds produced by growing colonies were absorbed into 0.2 ml of either 2% HCl or 10% citric acid (located as shown on Fig. 4Aa) for one week and analysed by (1) precipitation with Nessler's reagent and (2) HPLC chromatography on column (150 × 4 mm) of Watrex Cation 1-2, mobile phase citric acid 5 mM, dipicolinic acid 0.5 mM, conductomonitor III (TSP), performed by the Watrex company (Prague).

In the experiment presented in Figs 4 and 5, nitrogen content (N) in the HCl traps was measured in each sample as follows: 100 μl aliquots were mixed with 100 μl of 100 mM HCl and 200 μl of sodium phenolate (20% phenol, 8.9% NaOH). 600 μl of 1% sodium hypochlorite (water solution, containing 3% free chlorine) was added and the reaction mixture was heated at 80 °C for 1 h. After cooling to ambient temperature, the concentration of indophenol blue was measured at 625 nm. For calibrations, aqueous (NH₄)₂SO₄ solutions (1–20 μg N ml⁻¹) were used.

Induction of colony growth inhibition and turbid zone formation by NH₃ concentration gradients. Ammonia was generated in a small vessel containing NH₄Cl (70 mg) and 1 M NaOH (70 μl). The vessel was placed on the plate as shown on Fig. 4Ba, concurrently with 'giant' colony inoculation.

Received 30 June; accepted 2 September 1997.

- Shimkets, L. J. Social and developmental biology of the *Myxobacteria*. *Microbiol. Rev.* **54**, 473–501 (1990).
- Gross, J. D. Developmental decisions in *Dictyostelium discoideum*. *Microbiol. Rev.* **58**, 330–351 (1994).
- Ljungdahl, P. O., Gimeno, C. J., Styles, C. A. & Fink, G. R. SHR3: A novel component of the secretory pathway specifically required for localization of amino acid permeases in yeast. *Cell* **71**, 463–478 (1992).
- Kocková-Kratochvilová, A. *Yeasts and Yeast-like Organisms* (VCH, Weinheim, 1990).
- Serrano, R. in *The Molecular and Cellular Biology of the Yeast Saccharomyces* vol. 1 (eds Broach, J. R., Pringle, J. R. & Jones, E. W.) 523–585 (1991).
- Suomalainen, H. & Keränen, A. J. A. Keto acids formed by baker's yeasts. *J. Inst. Brew.* **73**, 477–484 (1967).
- Toda, T. *et al.* In yeast RAS proteins are controlling elements of adenylate cyclase. *Cell* **40**, 27–36 (1985).
- Goodman, L. E., Perou, Ch. M., Fujiyama, A. & Tamanoi, F. Structure and expression of yeast *DPR1*, a gene essential for the processing and intracellular localization of *ras* proteins. *Yeast* **4**, 271–281 (1988).
- Chan, R. K. & Otte, C. A. Isolation and genetic analysis of *Saccharomyces cerevisiae* mutants supersensitive to G1 arrest by a-factor and α-factor pheromones. *Mol. Cell. Biol.* **2**, 11–20 (1982).
- Mitchell, A. P. & Herskowitz, I. Activation of meiosis and sporulation by repression of the *RME1* product in yeast. *Nature* **319**, 738–742 (1986).
- Marini, A.-M., Vissers, S., Urrestarazu, A. & Andre, B. Cloning and expression of the *MEP1* gene encoding an ammonium transporter in *Saccharomyces cerevisiae*. *EMBO J.* **13**, 3456–3463 (1994).
- Schaap, P., Brandt, R. & van Es, S. Regulation of *Dictyostelium* adenylate cyclases by morphogen-induced modulation of cytosolic pH or Ca²⁺ levels. *Dev. Biol.* **168**, 179–188 (1995).
- Gimeno, C. J., Ljungdahl, P. O., Styles, C. A. & Fink, G. R. Unipolar cell divisions in the yeast *S. cerevisiae* lead to filamentous growth: regulation by starvation and RAS. *Cell* **68**, 1077–1090 (1992).
- Wright, R. M., Repine, T. & Repine, J. E. Reversible pseudohyphal growth in haploid *Saccharomyces cerevisiae* is an aerobic process. *Curr. Genet.* **23**, 388–391 (1993).
- Maráz, A. & Ferenczy, L. in *Protoplasts—Applications in Microbial Genetics* (ed. Peberdy, J. F.) 35–45 (University of Nottingham, Nottingham, 1979).

Acknowledgements. We thank V. Závada and J. Závada for comments, B. E. Griffin for help in improving this manuscript and I. Mocová for technical assistance. This work was supported by the Grant Agency of the Czech Republic, the Grant Agency of Charles University and by the Ministry of Education of the Czech Republic.

Correspondence and requests for materials should be addressed to Z.P. (e-mail: zdenap@prfdec.natur.cuni.cz).

corrections

Action-potential propagation gated by an axonal I_A-like K⁺ conductance in hippocampus

Dominique Debanne, Nathalie C. Guérineau, Beat H. Gähwiler & Scott M. Thompson

Nature **389**, 286–289 (1997)

Lüscher *et al.* have not shown that action potential propagation is normally reliable, as was inadvertently implied in our Letter. Their study demonstrates rather that action potentials can fail at axonal branch points in dorsal root ganglion cells, particularly during high frequencies of discharge.

- Lüscher, C., Streit, J., Quadroni, R. & Lüscher, H.-R. Action potential propagation through embryonic dorsal root ganglion cells in culture. I. Influence of the cell morphology on propagation properties. *J. Neurophysiol.* **72**, 622–633 (1994).

Activation of the transcription factor Gli1 and the Sonic hedgehog signalling pathway in skin tumours

N. Dahmane, J. Lee, P. Robins, P. Heller & A. Ruiz i Altaba

Nature **389**, 876–881 (1997)

The contents summary page in the issue of 23 October included an item on this Letter that was somewhat misleading. A more accurate summary reads: "Patients with basal cell nevus syndrome develop basal cell carcinomas (BCCs) early in life and carry mutations in the *Patched* gene, which encodes a receptor for the Sonic hedgehog ligand. These findings implicated the activation of the Sonic hedgehog signalling pathway in the familial or inherited form of BCC. However, the molecular mechanisms underlying the development of sporadic BCCs, the commonest form of skin cancer in fair-skinned adults with over a million cases a year worldwide, remained unknown. Now Dahmane *et al.* provide compelling evidence that virtually all sporadic BCCs have the Shh signalling pathway activated as determined by the expression of the zinc finger transcription factor Gli1, the final target and mediator of Shh signalling. The work predicts that any mutations that lead to the activation of this pathway in basal cells, and thus to Gli1 transcription and function, will cause basal cell cancer. Moreover, work in model organisms shows that inappropriate expression of Gli1 in the skin leads to the development of epidermal tumours. Gli1 may thus be both a marker and cause of BCC formation, making prospects for early diagnosis and possible treatment of this widespread type of skin cancer feasible."